貝氏主題混合資訊檢索模型

Bayesian Topic Mixture Model for Information Retrieval

吳孟淞 許軒睿 簡仁宗
國立成功大學資訊工程學系
Department of Computer Science and Information Engineering
National Cheng Kung University
mswu@chien.csie.ncku.edu.tw

摘要

在自動文件處理之相關研究中，我們常利用機率主題模型從字詞相互關係推斷並建立潛在主題變數。在機率潛在語意模型(PSLA)裡，文件中的每一個字詞在混合模型即視爲一個樣本，其混合成分是使用多項分佈來表示的。然而，多項分佈方式沒有考慮到文集中發生的突發現象。雖然 PLSA 模型可以顯示多重主題樣式，但是每個主題模型都十分簡單。在本研究中，我們提出一種新型之貝氏主題混合模型來解決多項分布固有的一些問題。使用 Dirichlet 分佈表示每一個主題的條件機率分佈，在相同種類內的不同的文件經由不同的多項分布來產生。在 TREC 文件集之資訊檢索實驗上，利用文件檢索及文件模組化之評估來驗證貝氏主題模型的優越性。

Abstract

In studies of automatic text processing, it is popular to apply the probabilistic topic model to infer word correlation through latent topic variables. Probabilistic latent semantic analysis (PLSA) is corresponding to such model that each word in a document is seen as a sample from a mixture model where mixture components are modeled by multinomial distribution. Although PLSA model deals with the issue of multiple topics, each topic model is quite simple and the word burstiness phenomenon is not taken into account. In this study, we present a new Bayesian topic mixture model (BTMM) to overcome the burstiness problem inherent in multinomial distribution. Accordingly, we use the Dirichlet distribution for representation of topic information beyond document level. Conceptually, the documents in the same class are generated by the associated multinomial distribution. In the experiments on TREC text corpus, we show the results of average precision and model perplexity to demonstrate the superiority of using proposed BTMM method.

關鍵詞：貝氏機率模型，圖形模型，機率潛在語意模型，Dirichlet 事前機率，資訊檢索

Keywords: Bayesian model, Graphical model, PLSA, Dirichlet Prior, Information Retrieval

一、緒論

隨著資訊大量氾濫，各種數位文件（digital documents）的遽增，使得資訊檢索精確度和文件模型的建立日顯重要。在資訊檢索和機器學習研究上，統計型本文模型(statistical text model)已逐漸成爲一個重要的議題。就資訊檢索的研究者而言，大多數將

文件視爲是 bag-of-word 的表示法，嘗試用統計的方法，擷取文字的特徵以建構資訊檢索的模式，此類方法亦稱爲向量空間模型[32]。Bag-of-word 的缺點是不考慮人類語言的同義字詞（synonym）以及多義字詞（polysemy）。再者，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。在文獻上，已有一些文件表示法被提出解決 bag-of-word 方面的一些問題。首先，潛在語意分析(Latent Semantic Analysis, LSA)[10]，是將文件以“字詞－文件”矩陣表示的方法。透過奇異值分解(Singular Value Decomposition, SVD)將文件投射到一個低維度的語意空間，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。在資訊檢索和語音辨識上已證明是有價值的分析工具[2][3][24]。第二，機率模型(Probabilistic Model)的基本假定爲觀測資料下的一個生成模型，此模型反應資料本身的架構。目前，已有一些機率模型的技術被廣泛地使用。例如，機率潛在語意分析(Probabilistic Latent Semantic Analysis)[16][17]以及 Latent Dirichlet Allocation[6]。PLSA 模型作法是擷取與文件關聯的意向模型(Aspect model)[18]。PLSA 模型有幾項缺點[6]，首先，是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。LDA[6]爲一個較完整的生成模型，其方法是將每一篇文件的機率視爲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。再者，文件以多項分佈表示法，無法有效取得字詞在文件中的突發現象(burstiness phenomenon)[12][25]。所謂「突發現象」意指，字詞在文件中出現過一次之後，很有可能會再出現的情形[22]。一般而言，字詞在文集裡一般分爲三種範疇，即常見(common)、一般(average)和稀有(rare)。雖然多項式表示能獲得常見字詞的突發性，但是對於一般和稀有字詞的突發性並未被正確的模組化。而透過 Dirichlet 分佈來替代多項分佈，可以趨緩突發現象的問題[25]。在本研究中，對於機率和主題混合模型問題感興趣，將探討幾個較先進的圖形模型[6][16][23][25]，期望藉由相關背景，來改善現有的文件模型架構。本文中以 PLSA 機率模型爲基礎，在混合模式的結合上，透過貝氏方法使用 Dirichlet 分佈決定各個分配所佔的比例，稱之爲貝氏主題混合模型(Bayesian Topic Mixture Model, BTMM)。透過 Gibbs 抽樣法來估計所需的參數。Gibbs 抽樣法的優勢是不需要明確地表示模型參數，可以在字詞分配到代表的潛在主題方面，簡單定義模型。本研究利用貝氏主題混合模型進行資訊檢索相關研究，所獲得成果對於改善搜尋系統檢索較易具有相當的應用價值。此外也可提供相關領域如資料探勘、機器學習等領域進行深入探討。本文接下來章節組織如下。第二章探討目前文獻中各種相關的文件模型研究方法。第三章將說明本文所提出的方法，並比較幾種主要模型的差異。第四章爲本文提出的方法和其他作法比較實驗效能分析的結果，用以證明本研究方法的效益及結果討論。最後，第五章爲本文的結論以及未來的研究方向。

二、相關文獻探討

在許多的應用上，資訊檢索和機器學習可以說密不可分。本章，我們將探索一些較具體、熟知的機率統計模型。首先，簡單描述在資訊檢索中較常見的文件表示法[10][32]。接著，針對廣泛的生成模型做更深入的探討，其中包含一些機率模型和混合模型等圖形模型表示式[6][16][17][23] [31]。

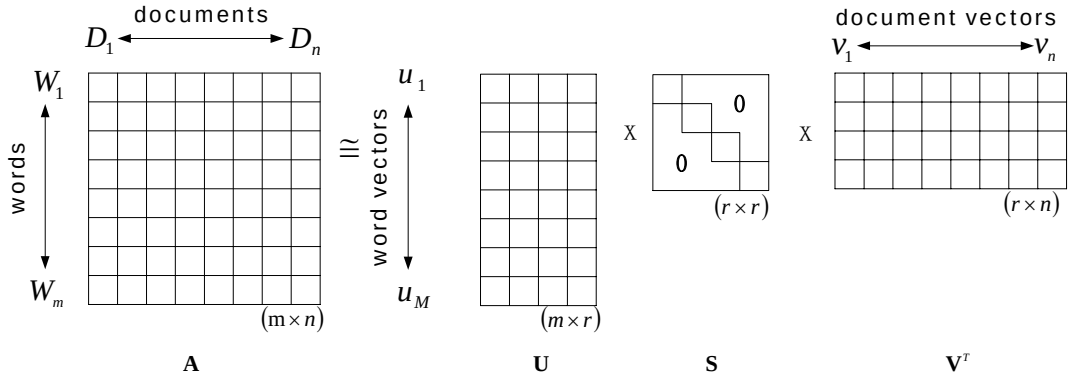
(一)、文件表示法

在資訊檢索系統中，文件通常由向量表示，意指視爲特徵的字詞出現在每篇文件的現象，此種表示法稱爲 bag-of-word 或者向量空間模型(Vector Space Model)[32]。其中

w_i 表示字典中的字詞在文件中出現的頻率值，而字典通常由文件集中的訓練集合所擷取得到。整個文件集可以透由字詞文件矩陣來表示，如下所示

$$\mathbf{A} = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1m} \\ w_{21} & w_{22} & \cdots & w_{2m} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \cdots & w_{nm} \end{bmatrix} \tag{1}$$

其中 w_{ij} 表示字典中的第 i 字詞在第 j 篇文件中出現的頻率值。在上述表示法中，缺乏任何有關字詞之間的語意訊息。因此，有其他學者考慮此類相關訊息來描述文件，稱爲潛在語意分析（Latent Semantic Analysis, LSA）[10]。LSA 基本的概念是以低維度的共同語意因子呈現原先文件和字詞之間的關聯。利用奇異值分解（Singular Value Decomposition, SVD）找出字詞對應文件的語意結構，可將高維度的矩陣資料降低爲 r 維度大小之特性。其奇異值分解之架構示意圖，如圖一所示。



圖一、奇異值分解之架構示意圖

(二)、文件混合模型之探討

1、Mixture of Unigrams

Mixture of Unigram (MU)模型是將 Unigram 模型經由離散隨機主題變數而擴增[31]。在此混合模型下，每份文件經由所選擇的主題所產生，接著，從主題相關的多項式獨立產生字詞。其文件的機率表示如下

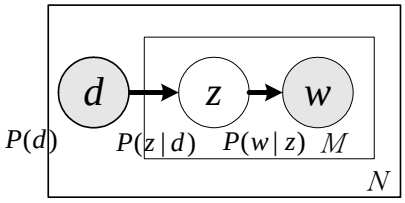
$$\begin{aligned} P(w) &= \sum_z P(w \mid z) P(z) \\ P(d) &= \prod_w P(w) = \prod_w \sum_z P(w \mid z) P(z) = \sum_z P(z) \prod_w P(w \mid z) \end{aligned} \tag{2}$$

當整個文集(corpus)被估計時，字詞分佈可以視爲在每一個文件對應一個主題的假設之下的主題表示。

2、Probabilistic Latent Semantic Analysis

先前所提 LSA 模型在文件和字詞上的呈現，並非以統計觀點出發。因此，Hofmann 提出機率潛在語意分析模型（Probabilistic Latent Semantic Analysis, PLSA）[16][17]，其

模型如圖二所示。PLSA 模型不同於 LSA 將文件和字詞向量投射至潛在語意空間的做法，其方法是以 Aspect Model 作爲主要架構[18]，使用機率密度函式作爲已觀察到的文件和字詞之間潛在語意關聯性的呈現方式，並利用最大相似度估測法則，結合了 EM 演算法[11]推估出隱含的模型參數。PLSA 模型目前被廣泛應用在多個領域，包括文件分段[7]、網頁探勘[19]、語音辨識技術及語言模型調適[1][8][9][29]等應用。



圖二、PLSA 模型示意圖

PLSA 模型主要的特徵，是針對字詞和文件共同事件尋求一個生成模型[16][17]。本文資料集是由字詞-文件對 (d,w) 所組成，文件以 $\mathbf{d} \in \{d_1,\ldots,d_N\}$ 表示，其個數爲 N；另外，字詞以 $\mathbf{w} \in \{w_1,\ldots,w_M\}$ 表示，字典相當於是 M 個字詞所形成之集合。假設每一字詞在給定的文件中潛在主題 $\mathbf{z} \in \{z_1,\ldots,z_K\}$ 下產生。將字詞-文件對 (d,w) 共同出現(co-occurrence)的聯合機率以式(3)表示

$$\begin{aligned} P(d,w) &= \sum_z P(z)P(w \mid z)P(d \mid z) \\ &= P(d)\sum_z P(w \mid z)P(z \mid d) \end{aligned} \tag{3}$$

在 PLSA 模型中，文件則經由 $P(w \mid z)$ 的因子的混合描繪其特性。將 z 視爲潛在變數，可以容易地對 PLSA 模型利用 EM 演算法來學習參數。最大化對數相似度可以表示成：

$$L_{\text{PLSA}} = \sum_d \sum_w n(d,w) \log P(d,w) = \sum_d \sum_w n(d,w) \sum_z P(z)P(d \mid z)P(w \mid z) \tag{4}$$

其 $n(d,w)$ 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

$$P_{\text{PLSA}}(z \mid d,w) = \frac{P(z)P(d \mid z)P(w \mid z)}{\sum_z P(z)P(d \mid z)P(w \mid z)} \tag{5}$$

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。其所有參數的更新如下

$$\hat{P}_{\text{PLSA}}(w \mid z) = \frac{\sum_d n(d,w)P(z \mid d,w)}{\sum_w \sum_d n(d,w)P(z \mid d,w)} \tag{6}$$

$$\hat{P}_{\text{PLSA}}(d \mid z) = \frac{\sum_w n(d,w)P(z \mid d,w)}{\sum_d \sum_w n(d,w)P(z \mid d,w)} \tag{7}$$

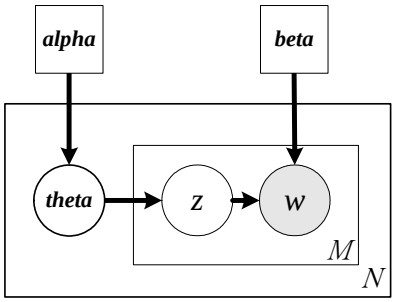
$$\hat{P}_{\text{PLSA}}(z) = \frac{\sum_d \sum_w n(d,w)P(z \mid d,w)}{\sum_d \sum_w n(d,w)} \tag{8}$$

PLSA 在資訊檢索中，可以藉由低維的”潛在”空間代替原始文件的表示。在 Hofmann[16][17]裡，以 $P(z \mid d)$ 作爲在低維空間之文件的組成，對於未看見(unseen)之文件或查詢句，經由最大化對數相似度和固定 $P(w \mid z)$ 及計算而得。

3、Latent Dirichlet Allocation

近幾年來，Latent Dirichlet Allocation (LDA)被提出來模組文集的潛在主題[6]。在大詞彙自動語音辨識系統下使用在語言模型的調整[30][33]，以及其他機器學習應用上皆有不錯的成效[4][5]。LDA 主要是克服 PLSA 模型中上述的缺點，比較 LDA 與 PLSA 模型相異之處，在於 LDA 將每一篇文件的機率都視爲潛在主題中隨機字詞機率的混合模型，藉此取得該篇文件出現的機率值。LDA 模型使用隨機變數 θ 來代替 PLSA 模型中 $P(z \mid d)$ 參數。θ 和 z 有相同的維度，表示文件中主題的混合。θ 對每一文件從 Dirichlet 分佈取樣，代替估計每一訓練文件的混合機率 $P(z \mid d)$ ，對 PLSA 模型而言，LDA 所需要的參數量較少。在 PLSA 模型中，有 $K*N$ 個 $P(z \mid d)$ 參數，而 LDA 模型，對文件的取樣，θ 只需 K 個參數。

在 LDA 模型裡，假設文件從潛在主題上隨機混合取樣，透過字詞上的分佈描繪每一主題的特性。在此模型中，文件爲觀察變數，視爲字詞的集合，$\mathbf{d} \in \{1,\ldots,M\}$ ，每一字詞取決於未觀察變數(也就是 topic) z，表示在 $\{1,\ldots,K\}$ 的可能值，並且 K 超參數(hyperparameter)必須被決定。在文件空間裡，LDA 模型存在未觀察變數，$\theta = (\theta_1,\ldots,\theta_K), \theta_k > 0$ 且 $\sum_k \theta_k = 1$。其模型如圖三所示，α 表示爲主題混合 θ 之 Dirichlet priori，而字詞機率透過 $K * M$ 矩陣 β 參數化，其中 $\beta = P(w \mid z)$ 。



圖三、LDA 模型示意圖

文件 d 和主題混合 θ 的聯合分佈爲

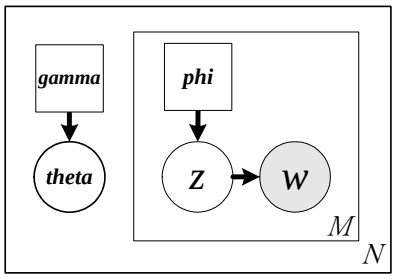
$$P_{\mathrm{LDA}}(\theta, \mathbf{d} \mid \alpha) = P(\theta \mid \alpha) \prod_w \left[\sum_z P(w \mid z) P(z \mid \theta) \right]^{n(d,w)} \tag{9}$$

其中，$n(d,w)$ 表示字詞 w 在文件 d 中出現的個數，$P(\theta \mid \alpha)$ 爲 θ 的 Dirichlet 機率分布。我們可以得到文件的邊際分佈

$$P_{\mathrm{LDA}}(\mathbf{d} \mid \alpha) = \int P(\theta \mid \alpha) \prod_w \left[\sum_z P(w \mid z) P(z \mid \theta) \right]^{n(d,w)} d\theta \tag{10}$$

然而，爲了估計這些參數必須計算事後分佈 $P(\theta, z \mid d)$ ，通常這些推論是不易實現的。在文獻上，一些基於變化方法的近似推論技術被提出，如 Variational Methods[6][21]、Expectation Propagation[28]和 Gibbs 抽樣法[14]。在此，對 Blei et al.[6]所提出的方法做說明，在圖形模型來說，Variational Method 是將一個複雜的圖形模型轉換爲一個簡單的

圖形模型，如圖四所示，期望簡化過後的模型能夠用正推論(exact inference)解決。



圖四、近似事後 LDA 模型之 Variational 分佈示意圖

Blei et al. [6] 定義一個分佈 $q(\theta, \mathbf{z} \mid \gamma, \phi)$ 的近似群，並且選擇 Variational Parameters γ 和 ϕ 接近真實的數值。Variational 分佈定義爲

$$q(\theta, z \mid \gamma, \phi) = q(\theta, \gamma) \prod_{z} q(z \mid \phi_z) \tag{11}$$

對於這新模型，可以經由 Variational Distribution 和 True Posterior 之間的 KL Divergence 最大化得到 $P(\theta, z \mid \mathbf{d}, \alpha, \beta)$ 的近似，

$$(\gamma^*, \phi^*) = \arg\min_{(\gamma, \phi)} D(q(\theta, z \mid \gamma, \phi) \| P(\theta, z \mid \mathbf{d}, \alpha, \beta)) \tag{12}$$

參數估測過程利用 variational EM，使得對數相似度最低界限(lower bound)最大化，基於近似事後分佈 $P(\theta, z \mid d)$ 的一種變化分佈來更新參數，透過下列兩個步驟迭代過程。在 E-step 中，使用變化的事後分佈近似，對每份文件找到多變參數 $\{\gamma, \phi\}$ 的最佳化值，

$$\phi_n \propto \beta \exp\{E[\log(\theta) \mid \gamma]\} \tag{13}$$

$$\gamma = \alpha + \sum_n \phi_n \tag{14}$$

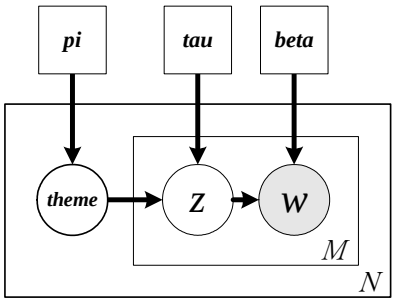
在 M-step 裡，使得有關模型參數對數相似度最小界限最大化，對條件多項參數的更新可以表示如下

$$\beta \propto \sum_d \sum_n \phi_{dn} w_{dn} \tag{15}$$

而參數 α 可以透過 Newton-Raphson 演算法求得[27]。Girolamin 和 Kaban [13]說明當 Dirichlet 分佈相同時，PLSA 模型實際上是 LDA 的一個特例。

4、Theme Topic Mixture Model

如前所述，LDA模型近似推論演算法並無法得到正解(Exact Solution)且計算複雜度增加。爲了克服這個問題，Keller和Bengio[23]提出一個正推論且易處理的模型，稱之爲 Theme Topic Mixture Model (TTMM)。在TTMM裡，文件空間的變數稱爲Theme，不同於LDA，TTMM對於topic的混合程度所佔的比例利用離散有限集(discrete finite set)來代替連續空間的使用。如圖五所示，此模型的觀察變數爲文件d，可視爲字詞w的集合，而未觀察變數爲theme，以 $\mathbf{h} \in \{1, \ldots, J\}$ 表示，以及topic，以 $\mathbf{z} \in \{1, \ldots, K\}$ 表示。其參數 π, τ 和 β 個別表示theme的混合程度所佔的比例 $P(h = j)$ 、topic給定theme的混合程度 $P(z \mid h = j)$ 以及每一字詞給定每一主題的機率值 $P(w \mid z)$ 。



圖五、TTMM 示意圖

每個文件可以視爲 theme h 的混合，表示爲

$$P(\mathbf{d}) = \sum_j P(h = j)P(d \mid h = j) = \sum_j P(h = j)\prod_w \left[\sum_z P(w \mid z)P(z \mid h = j)\right]^{n(d,w)} \tag{16}$$

其中，$P(d \mid h = j)$ 表示給定一個主題 $h = j$ ，其文件的生成機率，而 $n(d,w)$ 表示字詞在文件中的頻率，且 $\sum_w n(d,w) = n(d)$ 。假定文集 D 爲 N 篇文件的集合，給定文件模型，其文集 D 的對數相似度可以表示爲

$$L_{\text{TTMM}} = \sum_d \log\left[\sum_j P(h = j)\prod_w \left(\sum_z P(w \mid z)P(z \mid h = j)\right)^{n(d,w)}\right] \tag{17}$$

如同 PLSA 一樣，參數估計亦可經由 EM 演算法使得對數相似度最大化。在 E-step 中，潛在變數的事後機率被估計，如下所示

$$P(h = j \mid d) = \frac{P(h = j)\prod_w \left[\sum_z P(w \mid z)P(z \mid h = j)\right]^{n(d,w)}}{\sum_j P(h = j)\prod_w \left[\sum_z P(w \mid z)P(z \mid h = j)\right]^{n(d,w)}} \tag{18}$$

$$P(z \mid w, h = j) = \frac{P(z \mid h = j)P(w \mid z)}{\sum_{z'} P(z' \mid h = j)P(w \mid z')} \tag{19}$$

在 M-step 下，其對數相似度期望值是使用在上一階段估測的事後值，使得在標準化限制(normalization constraint)條件下最大化。模型參數的重新估測，可以表示爲

$$P_{\text{TTMM}}(h = j) = \frac{\sum_d P(h = j \mid d)}{\sum_{j'}\sum_d P(h = j' \mid d)} = \frac{\sum_d P(h = j \mid d)}{N} \tag{20}$$

給定條件限制 $\sum_{j'} P(h = j' \mid d) = 1$ ，可得

$$\hat{P}_{\text{TTMM}}(z \mid h = j) = \frac{\sum_d P(h = j \mid d)\sum_w n(d,w)P(z \mid w, h = j)}{\sum_d n(d)P(h = j \mid d)} \tag{21}$$

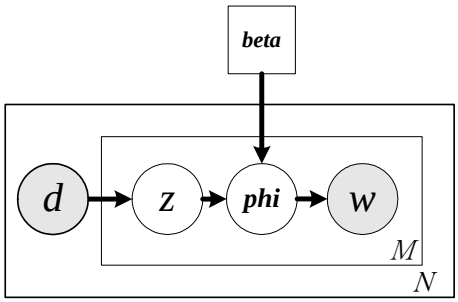
同理，給定條件限制$\sum_{z'} P(z' \mid w, h = j) = 1$，可得

$$\hat{P}_{\text{TTMM}}(w \mid z) = \frac{\sum_{d}\sum_{j} n(d,w)P(h = j \mid d)P(z \mid w, h = j)}{\sum_{w'}\sum_{d}\sum_{j} n(d,w')P(h = j \mid d)P(z \mid w', h = j)} \tag{22}$$

三、貝氏主題混合模型

(一) 模型定義

Madsen et al. [25]提到多項模型比較適合常見的字詞，但是對於其他較一般或是稀有的字詞無法有效獲得其突發現象。且多項式產生的計數分佈(counts distribution)基本上不同於自然本文的計數分佈。對於本文的模組化，Dirichlet 分佈在先前研究已被廣泛使用[6][12][25]。本文所提出之貝氏主題混合模型，同時擁有 PLSA 以及 Dirichlet 分佈的主要概念。使用 Dirichlet 模型於每一個主題的條件分佈，且文件裡每一個字詞可以由不同的主題所產生，使得在文件模型的表示上更豐富。模型架構如圖六所示



圖六 BTMM 示意圖

在 BTMM 裡，假設文件集 D 包含文件數 N 篇，文件表示爲 $\mathbf{d} \in \{d_1, \ldots, d_N\}$，而字典 V 相當於是 M 個字詞所形成的集合，字詞以 $\mathbf{w} \in \{w_1, \ldots, w_M\}$ 表示。未觀察變數爲主題，以 $\mathbf{z} \in \{z_1, \ldots, z_K\}$ 表示。假設文件 d 和字詞 w 條件獨立於給定的未觀察主題變數 z，對於所產生的模型參數，字詞是經由主題的多項分佈 ϕ 所產生，而對於字詞分佈的具體主題多項分佈 ϕ，可以從 Dirichlet priori 參數 β 對應的主題 z 得到。另外，文件是在 K 個潛在主題上使用 N 個混合數的多項分佈來表示，且 $\sum_{z} P(z \mid d) = 1$。在模型裡，參數集以集合 $\{\phi, \beta, P(z \mid d)\}$ 來表示，在推演過程中，使用 Dirichlet 分佈於主題多項分佈之上，因此隱藏參數 ϕ 可以被在外結合而不需要明確地被估計，此簡化過程，不需要在對 ϕ 取樣。如此一來，所需要的參數量共有 $KN + K$ 個。依據生成過程，字詞和主題的聯合分佈可以表示爲

$$P(w \mid z, \beta) = \int_{\phi} P(w \mid \phi)P(\phi \mid \beta, z)d\phi \tag{23}$$

而文件-字詞對 (d, w) 的聯合機率可以寫成

$$\begin{aligned} P(d, w \mid \beta) &= P(d)\sum_{z} P(z \mid d)P(w \mid z, \beta) \\ &= P(d)\sum_{z} P(z \mid d)\int_{\phi} P(w \mid \phi)P(\phi \mid \beta, z)d\phi \end{aligned} \tag{24}$$

(二) 推論

在我們的模型裡，潛在變數爲 $z_{d,w}$，即文件的字詞 w_d 所出現的主題。首先，觀察計算 $P(\mathbf{z}\mid\mathbf{w}_d)$ 的複雜度，此分佈和聯合分佈成正比。根據貝氏規則，對潛在變數 z 之條件事後分佈(conditional posterior distribution)給定爲

$$P(\mathbf{z}\mid\mathbf{w}_d)=\frac{P(\mathbf{w}_d,\mathbf{z})}{\sum_z P(\mathbf{w}_d,\mathbf{z})} \tag{25}$$

在此，我們以 $\{\mathbf{w}_d,\mathbf{z}\}$ 代表完整的觀察資料，$\mathbf{w}_d$ 表示文件向量。計算 $P(\mathbf{z}\mid\mathbf{w}_d)$ 意味著在大的離散狀態空間評估機率分佈。遺憾的是，這分佈無法直接被求得，主要是因爲分母部分加總比較難評估[14]，並且包含 Z^M 個離散隨機變數，其 M 表示在文件集中字詞的總數。就這觀點，在本研究中，嘗試以馬可夫蒙地卡羅法[20]以模擬參數的事後分佈(posterior distribution)來估計未知參數。如同文獻[14]一樣，使用 Gibbs sampling 來估計參數。對本研究提出的模型來說，Gibbs sampling 演算法容易實現，需要較少的儲存記憶體空間，並且在數度和執行上和現有的演算法具有競爭性。對於每一個字詞標記而言，Gibbs sampling 從對應的條件分佈中，給定其他字詞對於主題的分配，來估計目前字詞分配到主題的機率。然後目前字詞可以被分配到主題上，並且將此分配儲存下來，以便 Gibbs sampling 著手其他字詞計算時使用。爲了模擬 $P(\mathbf{z}\mid\mathbf{w}_d)$，Gibbs sampler 使用充分條件 $P(z\mid\mathbf{z}_{\neg i},\mathbf{w}_d)$ 執行 Markov chain。充分條件透過估算等式(25) 隱藏變數的方法可以寫成

$$P(z\mid\mathbf{z}_{\neg i},\mathbf{w}_d)=\frac{P(\mathbf{z},\mathbf{w}_d)}{\int_z P(\mathbf{z},\mathbf{w}_d)dz} \tag{26}$$

其中，$\mathbf{z}_{\neg i}$ 定義爲 $\mathbf{z}-\{z_i\}$，表示除了目前的字詞 w_i 之外，對所有字詞的主題分配。在 BTMM 中，聯合分佈可以被分解爲

$$P(z,w\mid d,\beta)=P(w\mid z,\beta)P(z\mid d) \tag{27}$$

等式右邊的兩個元素能夠被分別處理，第一項 $P(w\mid z,\beta)$ 可以由給定相關主題的被觀察字詞總數之多項式導出，如式(28)所示

$$\begin{aligned}P(w\mid z,\beta)&=\int_\phi P(w\mid\phi)P(\phi\mid\beta,z)d\phi=\int_\phi\frac{n!}{\prod_w n_w!}\frac{\Gamma(\sum_w\beta_w)}{\prod_w\Gamma(\beta_w)}\prod_w\phi_z^{n_z^{(w)}+\beta_w-1}d\phi\\&\cong\frac{\Gamma(\sum_w\beta_w)}{\Gamma(\sum_w\beta_w+n_z^{(w)})}\prod_w\frac{\Gamma(n_z^{(w)}+\beta_w)}{\Gamma(\beta_w)}\end{aligned} \tag{28}$$

其中，n_z 定義爲字詞 w 被分配到潛在主題變數 z 發生的次數。在式(28)中，$\prod_w\phi_w^{n_w}$ 和 $\prod_w\phi_w^{\beta_w-1}$ 結合是 Dirichlet 分佈 $P(\phi\mid n_w+\beta_w)$ 的未正規化變化形式，並利用 $\int P(\phi\mid n_w+\beta_w)d\phi=1$ 推導所得。過程中，不需導入參數 ϕ，因爲他們只是在被觀察資料 (d,w) 和對應主題 z 之馬可夫鏈的狀態變數之間的關聯統計。考慮式(28)中的分佈，只對包含索引 i 之潛在變數 z 乘積項保留，其他全部消去。更進一步地，利用等式

$\Gamma(x)=(x-1)\Gamma(x-1)$。因此，式(28)可以重寫爲

$$\hat{P}_{\text{BTMM}}(w\mid z,\mathbf{z}_{\neg i})=\frac{P(w\mid \mathbf{z})}{P(w\mid \mathbf{z}_{\neg i})}=\frac{\frac{\Gamma(n_z^{(w)}+\beta_w)}{\Gamma(\sum_{w'}n_z^{(w')}+\beta_w)}}{\frac{\Gamma(n_z^{(w)}-1+\beta_w)}{\Gamma([\sum_{w'}n_z^{(w')}+\beta_w]-1)}}=\frac{n_{z,\neg i}^{(w)}+\beta_w}{[\sum_{w'}n_z^{(w')}+\beta_{w'}]-1}\propto\frac{n_{\neg i,z}^{(w)}+\beta_w}{n_{\neg i,z}^{(\cdot)}+V\beta_{w'}} \tag{29}$$

同理，潛在主題分佈 $P(z\mid d)$ 可以被推得以下結果

$$\hat{P}_{\text{BTMM}}(z\mid d,\mathbf{z}_{\neg i})=\frac{n_{d,\neg i}^{(z)}}{\sum_{z'}n_{d,\neg i}^{(z')}} \tag{30}$$

其中，$n(d,w)$ 表示字詞 w 在文件 d 中出現的個數。最後，對於潛在變數，由式(29)、(30)我們可以推導出更新等式，其結果爲

$$P(z_i\mid \mathbf{z}_{\neg i},w,d)\propto P(w\mid z,\mathbf{z}_{\neg i})P(z\mid d,\mathbf{z}_{\neg i})\propto\frac{n_{\neg i,z}^{(w)}+\beta_z}{n_{\neg i,z}^{(\cdot)}+V\beta_z}\cdot\frac{n_{d,\neg i}^{(z)}}{\sum_{z'}n_{d,\neg i}^{(z')}} \tag{31}$$

其中，$n_{\neg i,z}^{(w)}$ 表示字詞 w 分配給主題 z 的次數 ，$n_{d,\neg i}^{(z)}$ 包含主題 z 在文件 d 裡被分配到一些字詞 w 的次數，而 $n_{\neg i,z}^{(\cdot)}$ 表示所有字詞分配給主題 z 的總數，標記 $\neg i$ 表示當前字詞 w_i 在這些計數已被移去，不被列入計算考慮。β表示 Dirichlet priori，在本模型裡，對全部字詞 β假設是相同的，亦即 β 的所有組成部分都相同。z_i 的初始被設定介於值 1 到 K 之間，決定馬可夫鏈(Markov chain)的初始狀態。然後執行幾個迭代次數，直到鏈接近目標分佈， z_i 目前值將會被記錄下來。

(三) 不同模型之關聯和比較

在本章節中，我們將討論並比較前面章節所描述的幾個模型。從主要的方程式看來，模型之間差異大同小異。爲了容易理解文件模型生成的差異。針對本文所提出的方法和第二章所提到的模型，如 PLSA、LDA 以及 TTMM 等，對其組成元素(字詞、主題及文件)之生成機率/分佈表示，簡單歸納如下表一所示。

表一、不同方法之各組成元素機率分佈表示

	Word	Topic	Document
PLSA	$P(w\mid z)$	$P(z\mid d)$	$P(d,w)$
LDA	$w\mid z,\beta\sim Mult(\beta)$	$z\sim Mult(\theta)$	$\theta\sim Dir(\alpha)$
TTMM	$w\mid z,\beta\sim Mult(\beta)$	$z\sim Mult(\tau)$	$h\sim Mult(\pi)$
BTMM	$w\sim Mult(\phi_z),\phi_z\sim Dir(\beta)$	$P(z\mid d)$	$P(d,w)$

假設在文件集裡有 N 篇文件，字典數大小爲 M，$|d|$表示文件長度，亦即在文件的字詞個數，K 爲主題(Topic)個數，J 爲 theme 數目以及群組個數爲 C。對於模型的空間複雜度比較，以表三做一簡單的闡述。各個模型所需的參數量，從表二可以得知，TTMM 需要 $J(1+K)+KM$ 個參數，而 LDA 只需 $K+KM$ 個參數。主要是由於連續分佈使用一

個參數，在 LDA 產生混合比例θ參數，取代在 TTMM 兩個離散分佈。除此，當文件透過主題(theme)被群聚在一起，如此 $J < N$，則 TTMM 的參數量可能少於 PLSA 的參數量 $KN + KM$。在 BTMM 模型中，字詞是經由主題 z 的多項分佈ϕ所產生，而對於字詞分佈的具體主題多項分佈ϕ，可以從 Dirichlet priori 參數β對應的主題 z 得到，其參數量比 PLSA 少，只需 $KN + K$ 個。

表二、對不同模型之空間複雜度比較

	PLSA	LDA	TTMM	BTMM
Parameters	$O(KN+KM)$	$O(K+KM)$	$O(J+JK+KM)$	$O(KN+K)$

四、實驗

(一)、實驗文集及設定說明

在本文的實驗中，我們使用 TREC 所收集的文集，分別爲 Associated Press newswire (AP) 88 和 Wall Street Journal (WSJ) 89，資料的統計資訊，如表三所示。我們所使用測試的查詢句子爲 Topics 101-150，主要取各個主題中的標題(title)和敘述(description)部分作爲查詢句，每個查詢句的平均長度爲 14.48 個字。文件會先經過 stop word 和 stemming 的前處理。本文分別對此兩文集以文件檢索和文件模組化驗證本文方法的正確性和可行性。在實驗中主要是針對 Language Model (LM)、PLSA、LDA 及本文所提出的 BTMM 做比較。對於潛在變數 k 的個數，初始實驗設定爲 16。實驗分爲兩個部分，第一評估各個模型應用在文件檢索上的效能，以 Precision-Recall curve 和 mAP 作爲評估的準則 [15]。第二個是以 perplexity 評估文件模型的效果。

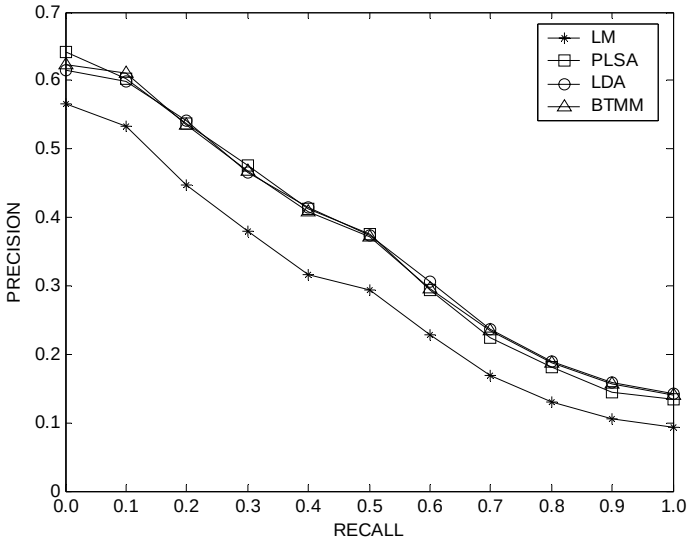
表三、TREC 文集的統計資訊

Collection	Description	Size (MB)	#Doc.	Vocabulary Size
WSJ89	Wall Street Journal (1989), Disk2	36.5	12,380	17,732
AP88	Associate Press (1988), Disk1	237	79,908	8,783

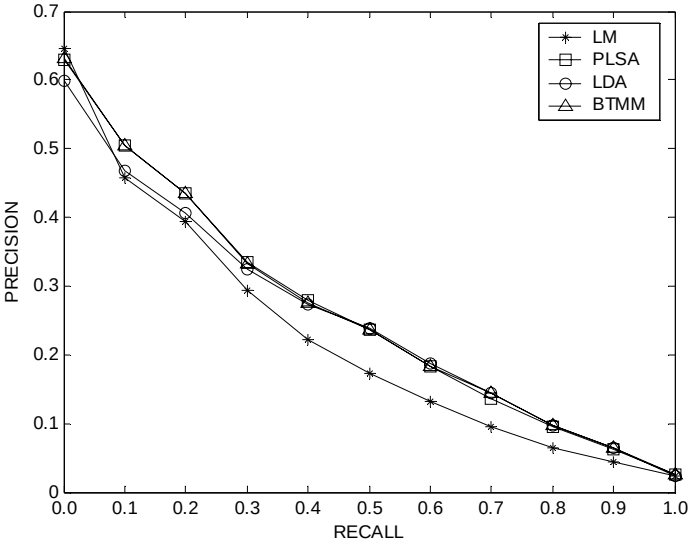
(二)、實驗結果

1、不同模型在檢索效能的影響

首先，比較不同的方法對 TREC 文件集在文件檢索上效能的比較。從圖七和圖八表示不同模型之 Precision-Recall 曲線，分別 WST89 和 AP88 的結果，而表四爲 mAP 在不同模型所計算的結果。從這些圖表當中，可以看出以主題爲基礎的文件模型，皆比語言模型有更好的效能。BTMM 的效能雖然比 PLSA 好，然而效果並不明顯。分析其原因，其影響的因素可能來自潛在主題變數 k 值的設定和參數值初始的設定。另外，文件前處理 stemming 亦可能造成影響。因此，在未來的實驗，將針對這些部分更進一步探討。



圖七、Precision-recall curves 對不同方法在 WSJ89 文集上的比較



圖八、Precision-recall curves 對不同方法在 AP88 文集上的比較

表四、LM、PLSA、LDA 以及 BTMM 在不同文集中 mAP 之比較

	LM	PLSA	LDA	BTMM
AP88	0.2128	0.2507	0.2411	**0.2536**
WSJ89	0.2761	0.3448	**0.3507**	0.3486

2、不同模型在文件模組化的評估

在文件模組化的實驗過程裡，以 WSJ89 爲實驗資料，將文件分爲兩個部分，三分之二的資料量作爲基礎模型的訓練資料集，共 7,931 篇文件，另外，三分之一部份做測試的文件資料集合，包含 4,449 篇文件。初步實驗結果如表五所示。從表中可以看出 BTMM 比 LM 和 PLSA 模型有較好的結果，其 perplexity 分別由 257.59 和 251.8 降至 250.42。

表五、不同模型之間 perplexity 之比較

	LM	PLSA	LDA	BTMM
Perplexity	257.59	251.8	248.63	250.42

BTMM 主要是改進 PLSA 中，字詞和主題之間的表示型態，以 Dirichlet 分佈替代原始的多項分佈，在字詞的主題分佈上導入 Dirichlet 事前機率，使得資訊更完整和豐富。然而，從實驗結果我們可以發現比 LDA 略差。針對此部分，我們將對字典個數的影響更進一步的探討分析。其分析結果如表六所示。我們分別選取字典字數一萬、二萬及三萬字來做對照，潛在主題變數個數設定爲 8。

表六、不同字典個數對 perplexity 值的影響

	10,000	20,000	30,000
LM	247	380	511
PLSA	240	372	504
LDA	**205**	**365**	505
BTMM	232	369	**495**

從表六可以得知，當字典數增加時，模型針對文字發生機率的預測分支度越高，所以 perplexity 都呈現上升的趨勢。當字典大小約爲 3 萬字時，BTMM 的 perplexity 比 LDA 低。主要原因是因爲當我們過度對字典數做刪減時，突發現象對模型的影響變得輕微。而由於 LDA 模型對文件階層加入事前機率，使得估算文件的主題分佈時，較貼近真實的分佈情形。然而，在字典數較大時，從實驗數據，可以發現突發現象較爲顯著，使得在文件中較稀有但卻具有鑑別性的字詞對模型產生影響，由於 BTMM 模型對字詞的主題分佈導入 Dirichlet 事前分佈，使得在 perplexity 的評估上略比 LDA 佳。

五、結論

本文中主要是以機率模型爲基礎提出一個貝氏理論的文件模型，致力解決 bag-of-word 表示法的問題，並對現有模型做改進，以期達到更好的效能。其架構延伸原始 PLSA 模型的概念，對於一個主題的條件分佈以 Dirichlet 代替原有的多項分佈表示，在此稱之爲貝氏主題混合模型。文中利用 Gibbs 抽象法估計模型未知參數，此方法的優點是不需要明確地表達模型參數且實做上比較容易，對記憶體需求量也比較少。在主題混合模型中，雖然假設文件可由不同主題所產生，但文件與字詞彼此之間是獨立的。然而，在真實世界裡，文件之間通常是有關聯的。例如，在新聞的文件標題中，可以分爲主要主題和次要主題。在 Tam 和 Schultz[34]的研究中，以 Dirichlet Tree[26]代替 LDA 中 Dirichlet Prior，使得潛在主題可以表達更多關聯。在未來的研究方向，對於文件模型演算法，我們擬延伸至層級概念，將文件以少量的概念或是主題來呈現，使得模型更具有強健性。另外，目前文件的機率模型表示法，大致以 Unigram 爲主，如何結合 *n*-gram 語言模型，使得文件模型更具強健性，亦是未來研究工作。

參考文獻

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

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

[3] 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.

[4] D. M. Blei and J. D. Lafferty, “Correlated topic model”, *Advances in Neural Information Processing Systems* (*NIPS*), vol. 18, pp. 147-154, 2006.
[5] D. M. Blei and J. D. Lafferty, “Dynamic topic model”, *Proceedings of the 23rd International Conference on Machine Learning*, pp.113-120, 2006.
[6] D. M. Blei, A. Y. Ng and M. I. Jordan, “Latent Dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, no. 5, pp. 993-1022, 2003.
[7] T. Brants, F. Chen and I. Tsochantaridis, “Topic-based document segmentation with probabilistic latent semantic analysis”, *Proceedings of the Eleventh International Conference on Information and Knowledge Management*, pp. 211-218, 2002.
[8] J.-T. Chien, M.-S. Wu and C.-S. Wu, “Bayesian learning for latent semantic language”, *Proceedings of European Conference on Speech Communication and Technology*, pp. 25-28, 2005.
[9] J.-T. Chien, M.-S. Wu and H.-J. Peng, “On latent semantic language modeling and smoothing”, *Proceedings of International Conference on Spoken Language Processing*, vol. 2, pp. 1373-1376, 2004.
[10] 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.
[11] 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.
[12] C. Elkan, “Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution”, *Proceedings of the 23rd International Conference on Machine Learning*, pp. 289-296, 2006.
[13] M. Girolami and A. Kaban, “On an equivalence between PLSI and LDA”, *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 433-434, 2003.
[14] T. L. Griffiths and M. Steyvers, “Finding scientific topics”, *Proceedings of the National Academy of Science*, vol. 101, pp. 5228–5235, 2004.
[15] D. Harman, Overview of the Fourth Text Retrieval Conference. 1995. Available at http://trec.nist.gov/pubs/trec4/overvies.ps.gz
[16] T. Hofmann, “Probabilistic latent semantic analysis”, *Proceedings of the Fifteenth Annual Conference on Uncertainty in Artificial Intelligence*, pp. 289-296, 1999.
[17] T. Hofmann, “Unsupervised learning by probabilistic latent semantic analysis”, *Machine Learning*, vol. 42, no. 1, pp. 177–196, 2001.
[18] T. Hofmann, “Unsupervised learning from dyadic data”, *Advances in Neural Information Processing Systems*, vol. 11. MIT Press, 1999.
[19] X. Jin, Y. Zhou and B. Mobasher, “Web usage mining based on probabilistic latent semantic analysis”, *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 197-205, 2004.
[20] M. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambrige, MA, 1999.
[21] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Sail, “Introduction to variational methods for graphical models”, *Machine Learning*, vol. 37, pp. 183-233, 1999.
[22] S. M. Katz, “Distribution of content words and phrases in text and language modeling”, Natural Language Engineering, vol. 2, pp. 15-59, 1996.
[23] M. Keller and S. Bengio, “Theme topic mixture model: A graphical model for document representation”, *in PASCAL Workshop on Learning Methods for Text Understanding and Mining*, 2004.
[24] 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.
[25] R. Madsen, D. Kauchak, and C. Elkan, "Modeling word burstiness using the Dirichlet distribution", *Proceedings of the 22nd International Conference on Machine Learning*, pp. 545-552, 2005.
[26] T. Minka, "The Dirichlet-tree distribution", in http://research.microsoft.com/~minka/papers/dirichlet/minka-dirtree.pdf
[27] T. Minka, "Estimating a Dirichlet distribution", *Technical Report, MIT*, 2000.
[28] T. Minka and J. Lafferty, "Expectation-propagation for the generative aspect model", *Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence*, pp. 352-359, 2002.
[29] D. Mrva and P. C. Woodland, "A PLSA-based Language Model for Conversational Telephone Speech", *Proceedings of International Conference on Spoken Language Processing*, pp. 2257-2260, 2004.
[30] D. Mrva and P. C. Woodland, "Unsupervised language model adaptation for mandarin broadcast conversation transcription", *Proceedings of International Conference on Spoken Language Processing*, pp. 1961-1964, 2004.
[31] K. Nigam, A. K. McCallum, S. Thrun and T. Mitchell, "Text classification from labeled and unlabeled documents using EM", *Machine Learning*, vol. 39, no. 2-3, pp. 103-134, 2000.
[32] G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*, New York: McGraw-Hill, 1983.
[33] Y.-C. Tam and T. Schultz, "Dynamic language model adaptation using variational Bayes inference", *Proceedings of European Conference on Speech Communication and Technology*, pp. 5-8, 2005.
[34] Y.-C. Tam and T. Schultz, "Correlated latent semantic model for unsupervised LM adaptation", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, pp. 41-44, 2007.