國立交通大學

資訊科學與工程研究所

碩 士 論 文

基於 Constrained-PLSA 之半監督式文件分群

Document Clustering with Labeled and Unlabeled Data Using Constrained-PLSA

研 究 生：陳俊憲

指導教授：李嘉晃 教授

中 華 民 國 一 百 年 六 月

基於 Constrained-PLSA 之半監督式文件分群
Document Clustering with Labeled and Unlabeled Data Using Constrained-PLSA

研 究 生：陳俊憲　　Student：Chun-Hsien Chen

指導教授：李嘉晃　　Advisor：Chia-Hoang Lee

國 立 交 通 大 學

資 訊 科 學 與 工 程 研 究 所

碩 士 論 文

A Thesis

Submitted to Institute of Computer Science and Engineering

College of Computer Science

National Chiao Tung University

in partial Fulfillment of the Requirements

for the Degree of

Master

in

Computer Science

June 2011

Hsinchu, Taiwan, Republic of China

中華民國一百年六月

基 於 Constrained-PLSA 之
半 監 督 式 文 件 分 群

學生：陳俊憲　　指導教授：李嘉晃 教授

國立交通大學資訊學院 資訊科學與工程研究所碩士班

摘要

目前網路上的資料相當龐大，可輕易取得非常多未標記資料；然而監督式學習方法，需要給足夠標記的資料做訓練分類模型，資料標記往往需要浪費大量人力以及時間；而非監督式學習方法雖然不需要標記資料，但是往往使用者在分群之前已經有些背景知識，理論上這些知識應該加入系統，讓系統可快速有效的分群，所以本論文加入少許標記的資料，利用這已知的資訊，來達到更好的效果，同時不用介入過多的人力來幫助資料的分群。本論文提出 Constrained-PLSA，這是一種半監督式學習的演算法，將些許標記資訊整合加入 Constrained-PLSA 演算法中，利用標記的資訊引導未標記的資訊導向正確的方向，使分群效果提升。最後實驗結果顯示只要些許的標記資料可以讓 Constrained-PLSA 達到穩定且不錯的效果。另外本論文也用 Constrained-PLSA 探討標籤分析，利用論文資料集做實驗，此資料集每篇文章包含了摘要和標籤兩個資訊，標籤是由使用者看完文章後所給定的關鍵字，因此標籤是一個很重要訊息；本論文分析出四種摘要和標籤的組合方式：Words only、Tags only、Words+Tags 和 Tags as words，利用這幾種組合方式做實驗，並用不同的分群演算法來討論分析哪個組合方式下，能使標籤有最好效能提升效果，在此實驗中也可看出 Constrained-PLSA 可以經由些許標記資料，有效提升分群效能。

Document Clustering with Labeled and Unlabeled Data Using Constrained-PLSA

Student：Chun-Hsien Chen　　Advisor：Prof. Chia-Hoang Lee

Institute of Computer Science and Engineering
College of Computer Science
National Chiao Tung University

Abstract

Text classification is of great practical importance today given the massive volume of online text available. Supervised learning is one of the popular techniques for tackling text classification problems. However, enough labeled data is necessary for supervised learning methods. Labeling must typically be done manually and it is a time-consuming process obviously. In general, unlabeled data may be relatively easy to collect. Although unsupervised learning method doesn't need any labeled data. But users often have some background knowledge before clustering. Practically, background knowledge should be included into algorithms to improve clustering accuracy. This paper extends PLSA clustering model to propose a Constrained-PLSA method, which is a semi-supervised learning algorithm. The Constrained-PLSA assumes that data is generated by a mixture model and the correspondence between each document and class label is one to one. By introducing the seeding documents as constraints, we show that Constrained-PLSA can estimate maximum likelihood in latent variable models using the Expectation Maximization (EM) algorithm. Experimental results show that Constrained-PLSA with a small amount of examples can effectively improve the performance. In addition, this paper also discusses tag usage using Constrained-PLSA. Academic paper data set is employed in this paper. Each paper consists of abstract and tag information. Tag is given by users after reading the article. This paper analyzes four combinations of abstracts and tags: "words only", "tags only", "words + tags" and "tags as words". The best one is presented in this paper. Meanwhile, the experimental result shows that Constrained-PLSA outperforms other clustering algorithms.

誌謝

本論文可以完成，首先要感謝的就是我的指導教授李嘉晃教授。有了教授的指引，我在研究的過程中不會手足無措；也謝謝教授的耐心指導，讓我對自然語言處理這個領域有更深的認識。我從教授的身上學到了做研究的方法，未來將成為我工作上的助力。接著要感謝三位辛苦的口試委員，謝謝教授們的建議，讓本論文的內容可以更加完整。

同時，我亦感謝這兩年來陪伴在我身邊的實驗室同學們、學長以及學弟。尤其是我的同學們，士元、智愷、而益，總是不斷的鼓勵我，對我的幫助更是多不勝數。兩年的時間，雖然不是很長，但是曾經有過的歡笑淚水，這些回憶會一輩子永存在我的心中。

最後，我要感謝我的家人，感謝你們對我的愛護和包容。謝謝你們在背後默默的支持，使我能夠順利的完成碩士學位。

心中有太多的感謝不知道如何表達，在此僅以本篇論文表示我對你們最誠摯的感謝，並祝福你們身體健康、萬事如意，謝謝。

陳俊憲　謹誌

資訊科學與工程研究所

智慧型系統實驗室

中華民國一百年七月

目錄

表目錄

圖目錄

第一章、緒論

1.1 研究動機

目前網路快速發展，網路上可得到相當多的資訊，如何讓使用者快速和正確的得到所需的資訊，成為一項重要的研究議題；要對龐大的資料集做整理、分析，已經不能以人工的方式來處理，因為這是一件複雜、又花費人力的事情。目前有許多領域也都在研究這些方面的問題，如資料探勘(Data Mining)、資料擷取(Information Retrieval)和自然語言處理(Natural Language Processing)等等，不管是對文字、圖片、聲音或者影像的處理，都是希望利用電腦來自動化幫忙處理。在這方面研究，我們需要一些快速又便利的方法，讓機器來幫我們整理、處理這些資料，讓使用者不用花費大量的人力介入，節省使用者時間，提升效率。

目前自動文件分類或分群的技術已經成功的應用於實務問題上，監督式學習(Supervised Learning)方法是一種熱門的文件分類方法，已經有許多的研究已經提出[1][2][3]。雖然現在網路能獲得的文件相當多，但是這些文件幾乎都是未標記(Unlabeled)的資訊，而監督式學習(Supervised Learning)方法需要給足夠的標記(Labeled)資料訓練分類模型，而標記資訊往往需要花費大量的人力去標記；而非監督式學習(Unsupervised Learning)方法雖然不需要任何標記(Labeled)資料的輔助，但實際情況下，我們往往會有些許分群資訊，或我們會知道一點背景知識(Background Knowlege)，因此如何將這些已知資訊或背景知識加入分群中，實務上是一項有效且有用的研究。本論文提出一個半監督式學習(Semi-Supervised Learning)，用已知之些許標記(Labeled)資料來輔助分群，不僅可以提高分群效果，同時不用花費大量人力去標記文章。

另外，現在網路上有許多的論壇網站，這些論壇網站提供使用者發表一些文章、提出一些自己的評論，或對文章給予一些標記資訊；其中一個論壇網站CiteULike是以學術論文為主的資料網，此網站中提供許多的資訊，主要有論文

的內容，還有使用者針對這些文章所給的標籤，我們抓下此論壇的文章資訊，利用這些資訊來進行文章的分群，本論文同時也探討如何利用標籤輔助論文文章分群，使分群的結果幫助研究者找到相關的文章。

1.2 研究目的

在大量的文章中，要如何對這些文章做分析，通常透過文章中出現的字來做處理，如果某兩篇文章中出現的字都很類似，而且這些字出現的頻率又很高，則這兩篇文章的相似度會相當高，這兩篇文章歸為同一個群的機率會很高。在資訊擷取、自然語言中，需要將文章轉換成讓電腦便於分析的格式，向量模型空間(Vector Space Model)是常用的一種方式，特徵為所有文章所出現的字，每篇文章則可以依照字出現的頻率(Term Frequency)來表示成一個向量，該表示法可以便於以數學來分析這些文章。

本論文以非監督式學習(Unsupervised Learning)分群分法 Probabilistic latent semantic analysis(PLSA)為基礎，此方法為非監督式方法，不需標記(Labeled)的資訊；本論文延伸 PLSA 演算法，提出一個半監督式學習(Semi-Supervised Learning)方法；將背景知識(Background Knowledge)有效加入到其演算法中，並研究如何讓標記(Labeled)資料幫助未標記(Unlabeled)資料導向正確的方向，使分群的效果提升。

另外在論文資料集中，有摘要和標籤兩個資訊，摘要是文章的內容大概簡介，標籤是使用者看過文章後，給予這篇論文的關鍵字或屬性，所以通常標籤的資訊是重要的。本論文研究如何將標籤資訊加入向量中，設計摘要和標籤的向量組合方式，使分群的效果達到最佳的情況。

1.3 論文架構

第一章：緒論，描述本論文之動機與目的。

第二章：相關研究，描述本論文會用到的演算法及技術原理。

第三章：Constrained-PLSA，介紹本論文所提出的演算法。

第四章：實驗過程與結果討論，分析實驗結果。

第五章：結論與展望，將本論文做個總結並討論系統未來走向。

第二章、相關研究

2.1 Semi-Supervised Learning

半監督式學習(Semi-Supervised Learning)是一種使用標記(Labeled)和未標記(Unlabeled)資訊的機器學習方法,近年來很多學者開始研究半監督式學習,並且有許多的半監督式演算法已經被提出,包含了 co-training[4][5]、semi-supervised Naïve Bayes[6]、Transductive support vector machines(TSVMs)[7]、graph-based approaches[8][9]、clustering-based approaches[10][11][12][13]等等,這些方法已被用在不同的領域,如自然語言處理(Natural Languages Processing)[9][14][15]、圖形識別(Pattern Recognition)[16][17]和資料擷取(Information Retrieval)[6][18]。

雖然半監督式學習能夠從未標記(Unlabeled)資料取得一些資訊,但如果在模型假設不符合問題結構的情況下,未標記資料並沒辦法提供輔助,反而可能會導致不好的分類效果[19];其中一個原因是標記(Labeled)的資料太少,因此許多的半監督式學習方法會對於模型給予一些很強的假設。例如:Nigam 所提出的 Semi-supervised Naïve Bayes[6],其假設是(1)資料是由混合模型(Mixture Model)成生成,(2)每篇文章和標記的群別是一對一對應關係。Blum 和 Mitchell 所提出的 Co-training[4]是假設每個資料的描述可以被分成兩種不同的角度。Graph-based 方法會以圖(Graph)描述資料,每個點代表標記(Labeled)和未標記(Unlabeled)的資料,每個邊代表資料和資料間的權重或相似度,這樣一來,分群的問題可以轉變成條件限制式的最佳化問題,目的是去導出一個式子可以預測標記和未標記的資料群別。通常假設的條件式須考慮以下兩種情況:(1)預測的結果要跟標記的資料越近越好。(2)整個圖應該要平整(smooth)[19]。本論文提出的 Semi-supervised learning 方法,與 Semi-supervised Naïve Bayes 的假設是相同的。

Semi-supervised learning 方法可以進一步分類成 Semi-supervised

classification 和 Semi-supervised clustering，Semi-supervised classification 利用標記和未標記資料建立一個更準確的分類器；而 Semi-supervised clustering 利用少量的標記資料去導正未標記資料的分群效果。基本上，unsupervised learning 在分群的過程中不需要有標記資料的加入，其目的是要將一個物件分到一個群體中，如此一來，在同一個群體中的物件，他們相似度一定大於不同群體中的物件，因此許多的分群演算法，主要目標在於最小化成本函數(Cost Function)，讓物件和其所代表的群體達到效能最好的情況。例如，K-means 區域性的最小化物件和群中心的平方和距離，Spectral clustering[20][21]利用圖形來代表資料集，他的目標函數(Objective Function)就會轉變成找圖中最小切割(Graph Minimum Cut)的問題。除了上述的演算法；PLSA[22]很成功的利用統計模型來分析文章的處理，已經被廣泛應用於多個領域上，例如主題模型(Topic Modeling)和文件分群(Document Clustering)。PLSA 是一個生成模型(Generative Model)，它是基於混合分解(Mixture Decomposition)衍伸出潛在類別模型(Latent Class Model)；這裡值得注意的是 PLSA 有兩種模型，一個是 aspect 模型，另一個是 statistical clustering 模型[23][24]，本論文所提出的 Constrained-PLSA 是利用 statistical clustering model 而不是 aspect model。

雖然 unsupervised learning 的方法在分群文件時不需要用到標記的資料，但是若能提供適合的種子(Seeds)，會將分群的結果導向正確的方向[11]。很顯而易見的，在分群演算法中加入一些預知的資料，可以增進分群的效果，通常預知的資料可以當成分群時的限制條件，當分群程序結束時，必須要滿足這些限制條件。Wagstaff[10]提出一種 semi-supervised 的 K-Means 的變形，叫做 COP-KMeans，此方法利用一些先知知識來設置限制條件；有兩種限制條件，第一種是必須連結(must-link)，此條件代表兩個物件必須要被分在同一個群內，第二種是不可連結(cannot-link)，兩個物件必須在不同的群內，這些條件應用在分群的過程中，讓分群的最後結果能滿足這些條件。Basu[11]利用初始標記資料

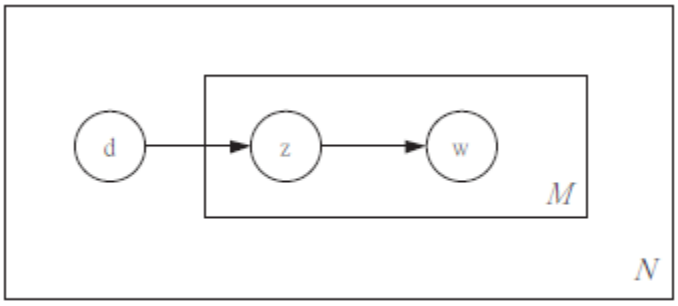
當作種子提出兩種 semi-supervised 的 K-Means 的變形，這兩種的名稱叫做 Seeded-KMeans 和 Constrained-KMeans，在 Seeded-KMeans 方面，種子只被用在 K-Means 初始化的時候，沒有被用在分群演算法中，在 Constrained-KMeans 方面，種子被用在初始化群中心和在分群運作的時候，保持標記資料的群組不改變，他們的實驗結果表示 Constrained-KMeans 比 Seeded-KMeans 效果好。標記資料在 Constrained-KMeans 和 Constrained-PLSA 的角色很相似，但分群方法是完全不一樣的;Constrained-PLSA 是 PLSA 的延伸，而 Constrained-KMeans 是 K-means 的延伸。

除了上面一段所提到的 K-means 變化方法，也有許多的 semi-supervised clustering 方法也是從其他的演算法延伸而來，例如，Finely 和 Joachims[26]提出一個 SVM 演算法是利用調整每個物件跟物件之間的相似度計算來訓練分群演算法，在實際問題中，所有的限制條件可能無法滿足，因此 Wang[27]提出一個有效的軟性限制(Soft-Constraint)演算法，讓限制條件盡可能滿足，此演算法能獲得一個較正確的分群結果。除了 Constrained-KMeans，其他與本論文提出的 semi-supervised clustering model 相關的研究，還有由 Zhou[13]所提出的文件分群，Topic-Sensitive PLSA，利用部分群組的限制條件來做 PLSA 的延伸修改，實際上，Constrained-PLSA 和 Topic-Sensitive PLSA 是不同的。Topic-Sensitive PLSA 會要求使用者選擇一些有興趣的文章當作他興趣的主題，這些文章符合使用者興趣，所以將這些文章當作興趣標記(Positive Labeled)文章，這個模型假設這些興趣標記文章屬於某一個主題 k，並假設這些文章為興趣(Positive)群體，如果其他文章不屬於主題 k 則代表這些文章為沒興趣文章，被定義為歸屬在非興趣(Negative)群體，Topic-Sensitive PLSA 的目的是要利用興趣標記文章和未標記(Unlabeled)文章，將未標記文章分類出有興趣和沒興趣的文章，如果未標記的文章被分到有興趣的群則代表這些文章是使用者有興趣的，換句話說，這實驗目的與分成兩群相似，利用興趣標記資料去導正分群結果。

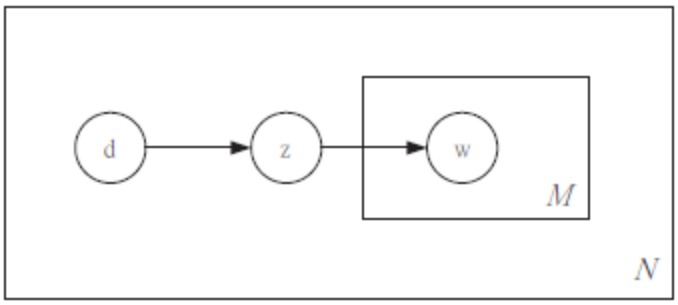
2.2 PLSA Model

從資料中學習一直是很熱門的研究領域，例如，機器學習(Machine Learning)、圖形識別(Pattern Recognition)和自然語言處理(Natural Language Processing)，都一直在研究如何有效的從資料中得到隱藏的資訊，甚至預測未來的結果，其目標為給定一個資料集，目的是從學習中建出一個模型，這樣一來可從模型中發現這些資料隱含的意義，除此之外模型還可以找到資料一些規律，然後就可以利用這些規率預測新的一筆資料的狀況，Hoffmann[24]應用 dynamic data 提出一個 unsupervised learning 的架構，dynamic data 指的是一個領域(domain)兩種物件的集合 $X=\{x_1,\ldots,x_n\}$、$Y=\{y_1,\ldots,y_m\}$， (x_i,y_j) 表示他們同時出現(co-occurrence)的資訊。dynamic data 被應用在很多的領域，像是文件分析(Text Analysis)、電腦視覺(Computer Vision)和計算機語言(Computational Linguistics)，在文件分析領域中，X 代表 document 的集合，Y 代表出現在 X 的 vocabulary 集合，co-occurrence 的資訊 (x_i,y_j) 則表示 y_j 這個字出現在文章 x_i 的次數。

根據上面的定義，latent semantic analysis(LSA)是一個分析文件(Document)和字(Word)相互關係的理論和方法，此方法提出了一些文章和字之間的概念，LSA 利用 singular value decomposition(SVD)來處理 document-term 矩陣，矩陣的低秩逼近(Low-rank Approximation)可以用來得知字和文章所包含的意義之間的關係。PLSA 為一延伸自 LSA 的統計模型，PLSA 有許多的特色，第一，PLSA 是一個 unsupervised learning 的方法，所以他不需要用到任何的標記資料，第二，PLSA 是一個生成模型，利用混合分解，來導出潛在類別模型，換句話說，只要系統擁有模型需要的參數就可以產生資料，第三，PLSA 的潛在變量(Latent Variable)可以觀察到很多的語意訊息(Semantic Information)，例如，PLSA 可以處理 Polysemy Problem，也就是說一個字能有許多意義的問題。



(a) PLSA aspect model



(b) PLSA statistical clustering model

圖2-1：two PLSA model

PLSA 有兩種 models，一個是 aspect model，另外一個是 statistical clustering model[23][24]，圖 2-1 中的(a)表示 PLSA aspect model，d 代表的是 document random variable，w 代表的是 word random variable，z 代表的是 latent variable，z 可以由每一組的(d_i, w_j)觀察取得，Z 是一個有限的集合$Z = \{z_1, \ldots, z_K\}$，換句話說就是要將資料分到 K 個群中，圖 2-1 中的(a)表示一篇文章可以連結到許多的潛在主題，每個主題又可以產生每個字的分布情形。

在 latent variable model 的情況下要計算 maximum likelihood 需要利用 Expectation Maximization(EM)[28]演算法，此演算中包含了兩個步驟，第一個為 E-step，第二個為 M-step，在 E-step 中是要利用 Bayes rule 求出 latent variables 的後驗機率(Posterior Probabilities)，以下為算出 latent variables 的後驗機率之等式。

$$P(z_k \mid d_i, w_j) = \frac{P(w_j \mid z_k)P(z_k \mid d_i)}{\sum_{l=1}^{K} P(w_j \mid z_l)P(z_l \mid d_i)}$$

由於 $P(d_i) \propto n(d_i)$，因此可以獨立提出來，在 M-step 中，要最大化 expected complete data log-likelihood，以下等式則是利用 E-step 算出的 posterior probabilities 來更新求解。

$$P(w_j \mid z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k \mid d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(z_k \mid d_i, w_m)}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k \mid d_i, w_j)}{n(d_i)}$$

第三章、Constrained-PLSA

此章節介紹本論文提出的演算法 Constrained-PLSA，此演算法是一個半監督式學習(Semi-Supervised Learning)方法。目前網路上有的資料相當龐大，可輕易取得，然而監督式學習方法，需要給足夠標記的資料做訓練分類模型，需浪費大量人力以及時間；而非監督式學習方法雖然不需要標記資料，但是往往使用者在分群之前已經有些背景知識，理論上這些知識應該加入系統，讓系統可快速有效的分群，所以本論文加入少許標記的資料，利用這已知的條件，來達到更好的效果，同時不用介入過多的人力來幫助資料的分群。

3.1 變數定義

以下會定義些符號，本節都是利用這些符號做推導。所有文章的集合為 D，共有 n 篇文章，則此集合的表示法為 $D=\{d_1,\ldots,d_N\}$，$d_1,\ldots,d_N$ 代表第 1 篇到第 n 篇的文章，本論文方法是要將每篇文章分群到他們屬於的群別，群別的集合為 C，共有 K 個群，C 的表示法為 $C=\{1,\ldots,K\}$。我們假設一個子集合的文章，此子集合的文章已知群別資訊，也就是我們的種子(Seeds)，表示法為 $d_i \in D^l$，已知的群別資訊屬於 C，$y_i \in C$，剩下的文章，為未知群別資訊的文章，這些文章的集合為 D^u，整個文章集合被分成兩份不相交的獨立集合，$D=D^l \cup D^u$。每篇文章可以表示成一個向量，$\langle w_{i,1},\ldots,w_{i,M}\rangle$，其中 $w_{i,j}$ 表示文字 w_j 出現在第 i 篇文章（d_i），V 代表所有文章出現的字之集合，$V=\langle w_1,\ldots,w_M\rangle$，$w_{i,j}$ 的值為 $n(d_i,w_j)$，也就是說文字 w_j 出現在文章 d_i 的次數。

3.2 Constrained-PLSA

PLSA 有兩種模型，一種是 aspect 模型，另一種是 statistical clustering 模型，aspect 模型是一篇文章對應多個 topic 之模型，而 statistical clustering 是一對一對應的，本論文採用的是 statistical clustering 模型，用此模型來延伸為 Constrained-PLSA。

本論文的模型架構假設與 Nigam[6]相同：第一個假設為資料產生的模型為混合模型(Mixture Model)，第二個假設為混合元件(Mixture Components)和群別為一對一對應關係。在這兩個假設之下，每篇文章利用混合模型來生成，用 Φ 來代表生成每篇文章的參數。首先依照混合權重(Mixture Weights)選擇一個混合元件。接著用所選擇的混合元件和其參數來生成一篇文章。因此文章 d_i 的 likelihood 由所有的混合元件的機率和所求得，如下方的方程式。

$$P(d_i \mid \Phi) = \sum_{k=1}^{K} P(z_k) P(d_i \mid z_k; \Phi) \tag{1}$$

群別的數量共有 K 個，z_k 代表第 k 個 component，因為每篇文章之間是獨立的情況，所以所有文章 D 的 log likelihood 可表示成每一文章 likelihood 之相乘，則 D 的 log likelihood 算法如下方程式。

$$\ln P(D \mid \Phi)$$

$$= \sum_{d_i \in D^l}^{N} \ln \sum_{k=1}^{K} P(y_i = z_k) P(d_i \mid y_i = z_k; \Phi) \tag{2}$$

$$+ \sum_{d_i \in D^u}^{N} \ln \sum_{k=1}^{K} P(z_k) P(d_i \mid z_k; \Phi)$$

基於每篇文章和群別是一對一對應(One-to-one Correspondence)，我們定義一個 K 維的 binary random variable π_i，$\pi_i = \langle \pi_{i1}, \ldots, \pi_{iK} \rangle$，當 $y_i = z_k$ 時，在這 K 維的向量中，其中有一個元素值 π_{ik} 為 1，其餘的值皆為 0；則 complete log

likelihood 可表示成下方程式。

$$\ln P(D|\Phi;\pi) = \sum_{d_i \in D}^{N} \sum_{k=1}^{K} \pi_{ik} \ln P(z_k) P(d_i | z_k;\Phi) \quad (3)$$

根據目前的模型架構，π_{ik} 取代了原本的期望值，利用 Jensen 不等式，會將方程式(3)complete log likelihood 限制小於 incomplete log likelihood 如方程式(2)所示，因此 EM 演算法可找到 locally maximum $\hat{\Phi}$。

$$Q = P(z_k | d_i) \quad (4)$$

$$\propto P(z_k) P(d_i | z_k) \quad (5)$$

$$\propto P(z_k) \prod_{j=1}^{M} P(w_j | z_k)^{n(d_i, w_j)} \quad (6)$$

$$= P(z_k) \prod_{j=1}^{M} \theta_{kj}^{n(d_i, w_j)} \quad (7)$$

$$= P(z_k) \exp\left(\sum_{j=1}^{M} n(d_i, w_j) \ln \theta_{kj}\right) \quad (8)$$

$$E[L^c] = \sum_{d_i \in D}^{N} \sum_{k=1}^{K} P(z_k | d_i) \ln P(z_k | \Phi) P(d_i | z_k;\Phi) \quad (9)$$

$$H = E[L^c] + \rho\left(1 - \sum_{k=1}^{K} P(z_k)\right) + \sum_{k=1}^{K} \tau_k \left(1 - \sum_{j=1}^{M} P(w_j | z_k)\right) = 0 \quad (10)$$

在 E-step 是去計算隱藏語意的 posterior 機率分布，定義為 Q。Q 是一個大小為 $n \times K$ 的機率矩陣，每一行 q_{ik} 代表文章 d_i 在群別 k 的機率。方程式(4)到方程式(5)是利用 Bayes 推導而來，方程式(5)到方程式(6)是利用 multinomial model 推導而來，其中 multinomial mode 在 language model 也是一種 unigram model，方程式(6)忽略計算 multinomial coefficient。方程式(7)採用了 θ 取代了原本的變數，此變數代表 topic-word 的分布，θ 中每列 θ_k 代表一個群別，其

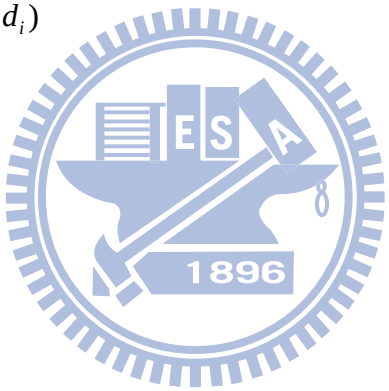
中一個元素θ_{kj}代表群別 k 產生字w_j的機率，方程式(8)是利用 natural logarithm 和 exponential function 將方程式(7)做轉換成另外一個形式。

因此可由方程式(9)complete log likelihood function 和機率的限制條件$\sum_{k=1}^{K} P(z_k)=1$、$\sum_{j=1}^{n} P(w_j \mid z_k)=1$來計算 Lagrange function。利用 Lagrange multipliers ρ和τ_k $(1 \le k \le K)$來得到方程式(10)的 objective function，利用方程式(11)和(12)probability mass function 來最大化 H。

$$P(z_k) \tag{11}$$

$$= \frac{\sum_{i=1}^{N} P(z_k \mid d_i)}{\sum_{k=1}^{K}\sum_{i=1}^{N} P(z_k \mid d_i)}$$

$$= \frac{\sum_{i=1}^{N} Q_{ik}}{\sum_{k'=1}^{K}\sum_{i=1}^{N} Q_{ik'}}$$



$$\theta_{kj} \tag{12}$$

$$= \frac{\sum_{i=1}^{N} P(z_k \mid d_i) n(d_i, w_j)}{\sum_{j=1}^{M}\sum_{i=1}^{N} P(z_k \mid d_i) n(d_i, w_j)}$$

$$= \frac{\sum_{i=1}^{N} Q_{ik} n(d_i, w_j)}{\sum_{j=1}^{M}\sum_{i=1}^{N} Q_{ik} n(d_i, w_j)}$$

3.3 Constrained-PLSA 演算法

Algorithm : Constrained-PLSA Algorithm

Input:

1. document-term matrix H of size N * M
2. K topics
3. $S_1..S_K$ 為 K 個 topic 挑出的 seed, $S_1..S_K \subset$ N documents

Output:

N 篇 document 的 K 個 topic 機率分布

Algorithm:

1. Initialize:

$$H_i = \frac{H_i}{\sum_j H_{ij}}, \text{for } i = 1,...,N \text{ 、}$$

$$\theta_k = \frac{1}{|S_k|}\sum_{d_i \in S_k} H_i, \text{for } k = 1,...,K \text{ 、c } = (\frac{1}{K}..\frac{1}{K})$$

2. Iterate:

(a)E-step

(1) $a_{ik} = \exp(\sum_j H_{ij}\log(\theta_{kj}))$

(2) $Q_{ik} = \frac{c_k a_{ik}}{\sum_{t=1}^{K} c_t a_{it}}$

(3) $Q_{S_1 1} = 1, Q_{S_2 2} = 1, ..., Q_{S_K K} = 1$

(4) $normalize(Q_{S_1}), normalize(Q_{S_2}), ..., normalize(Q_{S_K})$

(b)M-step

(1) $b_k = \sum_{i=1}^{N} Q_{ik} H_i$

(2) $\theta_k = \frac{b_k}{\sum_{j=1}^{M} b_{kj}}$

(3) $c_k = \frac{\sum_{i=1}^{N} Q_{ik}}{\sum_{t=1}^{K}\sum_{i=1}^{N} Q_{it}}$

(c)change = $\| Q - Q^{old} \|$

3. Terminate the iteration when change < t.

Input 的部分包含 N 篇文章的向量($H_1,...,H_N$)，向量的維度為 M，所以第 n 篇文章的向量可表示成 $H_n=(H_{n1},...,H_{nM})$，所有文章可表示成 N*M 的矩陣(H)，再來要給這些文章分成幾個群別(K)，最後要從這 I 篇文章中挑出一些文章來當作種子，每個群別都要有種子（ $S_1,...,S_K$ ），$S_1,...,S_K$ 集合內的數量依照群別的大小來定義，假設第 n 個群別有 1000 篇，取 1%的情況下，則取出 10 篇文章來當作種子，則 $|S_n|=10$ 。

Output 的結果為每篇文章在 K 個群別的機率分布，最後結果取機率最大的群別當作一篇文章所屬的群別。

Initialize(初始化)的部分，將每篇文章做 Normalize 的處理，H 的每列代表一篇文章，所以將 H 的每列做 Normalize，再來就是初始群中心的部分，取每個群別的種子($S_1,...,S_K$)的中心代表群別的群中心($\theta_1,...,\theta_K$)，$\theta_1,...,\theta_K$ 皆為實數 M 維的向量，每個值都不為負數，做 Normalize 處理使其總和為 1，θ_n 又可看成第 n 個群別出現所有 M 個字的機率分布向量，可從中得知群別最有可能出現哪些字，最後 c 代表每個群別發生的機率，預設給其平均值，表示每個群別發生的機率皆相同。

Iterate(迴圈)的部分，總共分成四個階段，(a)E-Step 求期望值並且修正 seeds 的權重、(b)M-Step 在做最大化更新，最後(c)階段在計算上回合和此回合變化量，判斷是否小於門檻值，停止迴圈。

(a)E-Step

主要在估計每篇文章在每個群別的機率分布，取每篇文章來對群中心算 multinomial distribution，計算公式的推導如下，也就是演算法中(a)的(1) a_{ik} 的部分。

$$P(H_i \mid \theta_k)$$

$$= \frac{(\sum_{j=1}^{M} H_{ij})!}{\prod_{j=1}^{M} H_{ij}!} \prod_{j=1}^{M} \theta_{kj}^{H_{ij}}$$

$$= \frac{(\sum_{j=1}^{M} H_{ij})!}{\prod_{j=1}^{M} H_{ij}!} \exp(\log(\prod_{j=1}^{M} \theta_{kj}^{H_{ij}}))$$

$$= \frac{(\sum_{j=1}^{M} H_{ij})!}{\prod_{j=1}^{M} H_{ij}!} \exp(\sum_{j=1}^{M} H_{ij} \log(\theta_{kj}))$$

$$= f(H_i) \exp(\sum_{j=1}^{M} h_{ij} \log(\theta_{kj}))$$

當 i 固定時，也就是計算第 i 篇文章時，不管 k 是多少的情況下，$f(H_i)$ 都是固定的，這邊可以將 $f(H_i)$ 去除，所以在演算法中沒有計算 $f(H_i)$ 這個值，演算法中(a)的(2) Q_{ik} 是在做 Normalize 的處理，Q_i 的總和還是 1，使其值還是一個機率分布的模型。

再來就是修正上述出來的結果，要修正種子所屬群別的權重，S_1 為第一個群別的集合，機率最大值為 1，因此將 S_1 內所有的文章的第一個群別設成 1，$S_1,\ldots,S_k$ 都要更新權重的處理，也就是(a)的(3)所看到的 $Q_{S_1 1}=1, Q_{S_2 2}=1,\ldots,Q_{S_K K}=1$。更新完權重後，其值不再是機率的模型，因此所有更改過值得要再做一次正規化的處理，也就(a)的(4)所處理的 $normalize(Q_{S_1}), normalize(Q_{S_2}),\ldots, normalize(Q_{S_K})$，讓數值再回到機率模型，使其總和為 1。

(b)M-Step

此階段主要利用上階段的結果，更新數值提供下個回合使用，(b)的(1)公式中 $b_k = \sum_{i=1}^{N} Q_{ik} H_i$，在計算新的群中心，利用每篇文章的每個群別機率，還有文章的向量來更新，第一個群別新的群中心算法就是將每篇文章第一個群別的機率乘上每篇文章的向量，依此，要分別計算出 K 個新的群中心，計算出來的結果並不是一個機率模型，所以(b)的(2)公式中 $\theta_k = \frac{b_k}{\sum_{j=1}^{M} b_{kj}}$ 要將其作 Normalize 處理，讓他回到機率模型，(b)的(3) $c_k = \frac{\sum_{i=1}^{N} Q_{ik}}{\sum_{t=1}^{K} \sum_{i=1}^{N} Q_{it}}$ 要計算新的每個群別的機率分布，統計第一個群別到第 K 個群別發生的機率。

(c)計算變化量

此階段要計算變化度，如同公式中 $\| Q - Q^{old} \|$，用新的每篇文章機率分布與舊的每篇文章機率分布做 1-norm 的處理，也就是計算每個元素的差值，此值要用來判斷是否達到收斂條件。

演算法最後判斷變化值是否小於 t，這邊 t 值預設為 10^{-3}，當上面(d)階段算出的結果小於 10^{-3} 時，則終止迴圈，代表變化量已經很小，達到收斂的狀況，最後回傳結果。

3.4 Constrained-PLSA 概念

此節介紹 Constrained-PLSA 的概念，我們已知背景知識(background knowledge)，也就是我們已知某些文章的群別資訊，利用這些資訊來修改 PLSA 演算法，讓原本是非監督式學習演算法的 PLSA，使其成為半監督式學習演算法，以下在介紹如何加入這些標記資料的想法，說明修改的兩個部分，為什麼如此修改。

1. 群中心修正

在非監督式學習演算法 PLSA 中，初始群中心的部分，是由所有的文章隨機挑取，如下圖 4-1 所示，三個圓圈為三個群別，圓圈內的點為文章，所以隨機挑取的情況，可能會挑到同一群別內的文章當作初始群中心，如圖中三角形的部分，都挑到相同群內的文章當作群中心。因為我們已知一些標記的資料，這些標記資料可以幫助我們找到較好的初始群中心。

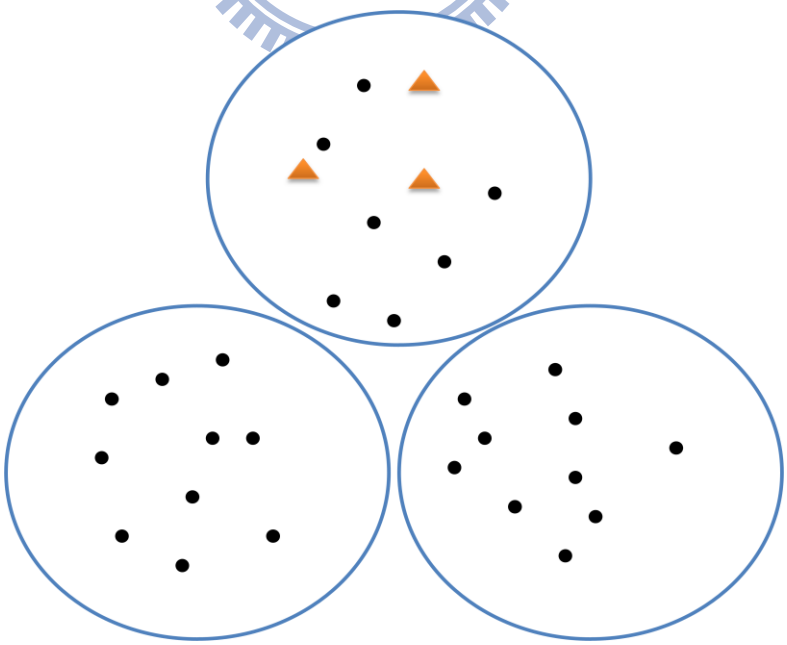


圖3-1：原始初始群中心

我們各群別取一些文章來當作我們的種子，如下圖 4-2 所示，紅點為三群內我們已知文章的群別，利用這個已知條件，我們即可來定義群中心位置，使分群效果提升。

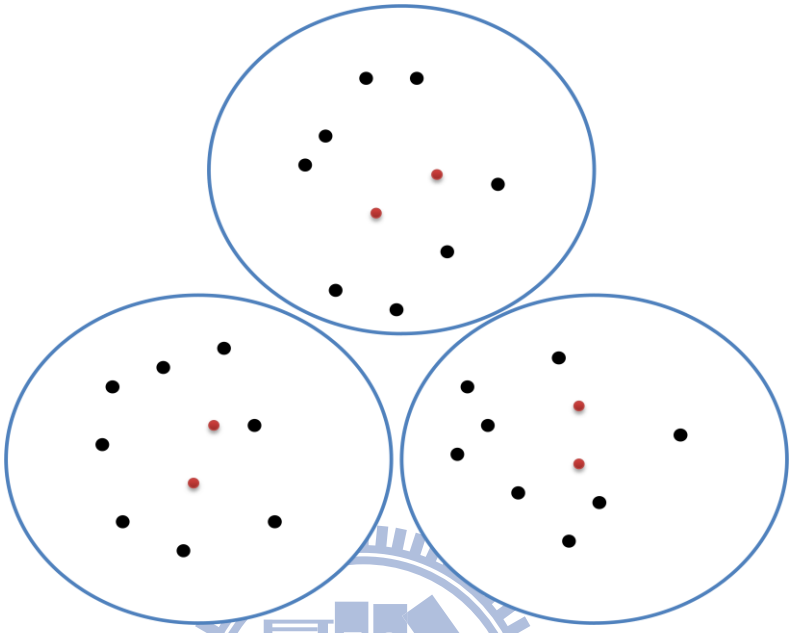


圖3-2：半監督式種子分布

我們利用這些種子的中心來代表群中心，如下圖4-3所示，三圓圈內的種子，我們取其中心當作我們的群中心，如圖中的三角形部分。

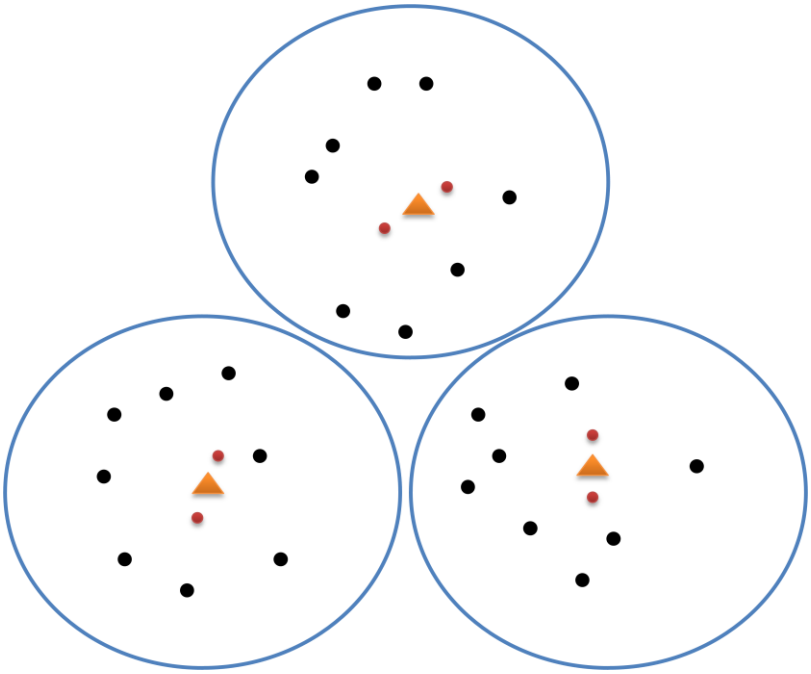


圖3-3：半監督式種子分布

2.修正種子所屬群別機率

由於我們已知種子的群別，所以要提高種子所屬群別的機率，如下圖 4-4 所示，我們已知紅色點屬於第一個群別，因為機率最大值是 1，所以將此點在第一群別的機率修正成 1，然後再做一次 Normalize，使其還是一個機率分布的模型，依此類推，每個種子都要修正他們所屬群別的機率。修正這些種子的機率，目的是為了新的群中心的計算，新的群中心的計算會用到每個文章的機率分布和每篇文章的向量，來調整新的群中心，所以將這些種子所屬的群別機率提高，有助於新的群中心導向正確的位置。

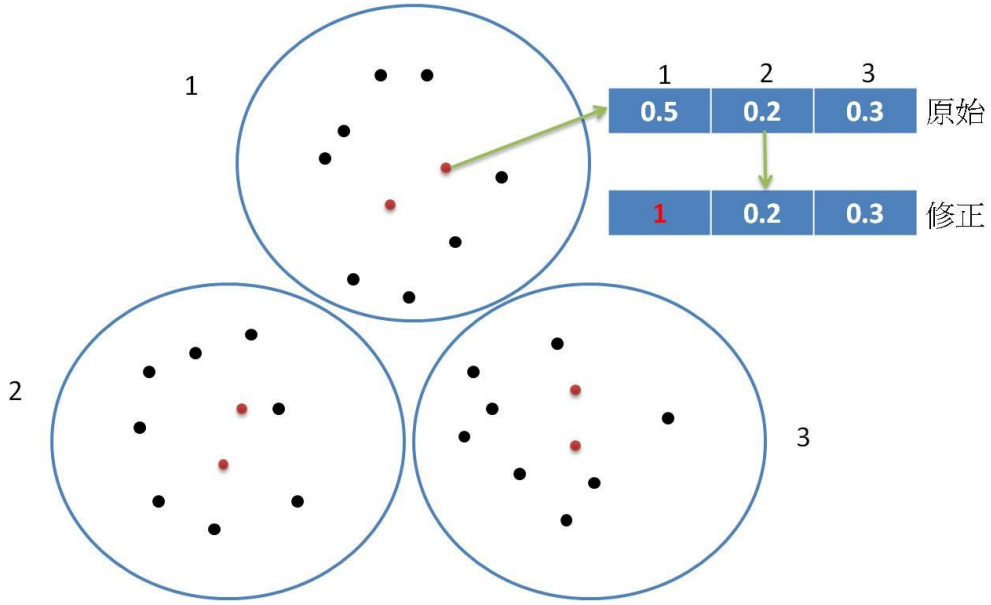


圖3-4：修正種子權重

第四章、實驗過程與結果討論

4.1 標籤分析

4.1.1 摘要標籤定義

此章節介紹標籤分析，以及如何利用論文資料集來做實驗。這論文資料集內的每篇文章包括論文的摘要、這篇論文所屬的類別和網路上使用者給這篇論文標籤；標籤通常是對這些文章分群的重要資訊，因為標籤通常代表文章的關鍵字，所以這部分是分析要如何使用標籤來輔助論文的分群，讓分群的效果達到最好的情況。

第五章實驗的部分利用了許多的方法來對這些向量組合方式進行實驗，分析哪種情況下的組合，使得多數方法會有明顯的效果提升，表示在這種組合的情況下，能夠達到最好的效果。

3.2 節在介紹論文資料集的取得方式，是從哪些網站收集而來，說明這些網站所提供的資源，要如何利用這些得知的資源進行查詢，最後是論文資料集的架構，說明此資料集的結構。

3.3 節在介紹摘要和標籤的組合方式，要怎麼利用這些組合表示一篇文章的向量，使得分群的效果提升。

4.1.2 論文資料集收集

本論文利用 Microsoft Academic Search[1]提供的論文資訊進行資料的收集，在這個網站中有列出每個類別下重要的期刊資訊，如下圖 4-1 所示，圖中 1.是代表類別 Databases、2.是代表在 Databases 下的期刊名稱。

1 http://academic.research.microsoft.com/

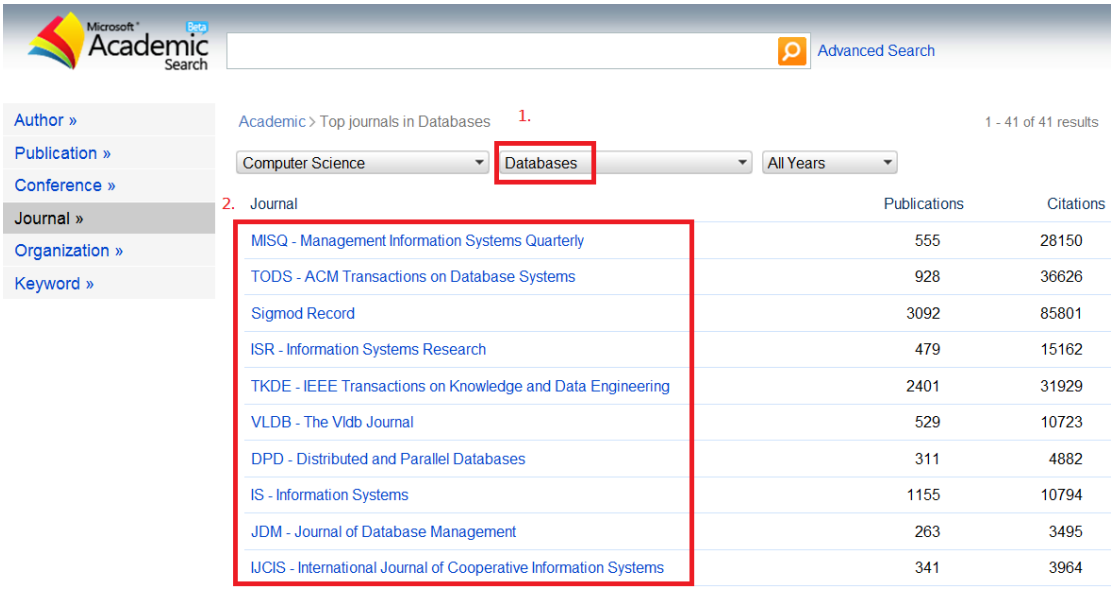


圖4-1：Microsoft Academic Search

從 Microsoft Academic Search 取得每個類別的期刊名稱後，到 CiteULike[2] 論文網站進行搜尋，搜尋這些期刊的論文清單，如下圖 4-2 所示，利用關鍵字 journal 搜尋 Databases 類別下的“IEEE Transactions on Knowledge and Data Engineering”期刊。

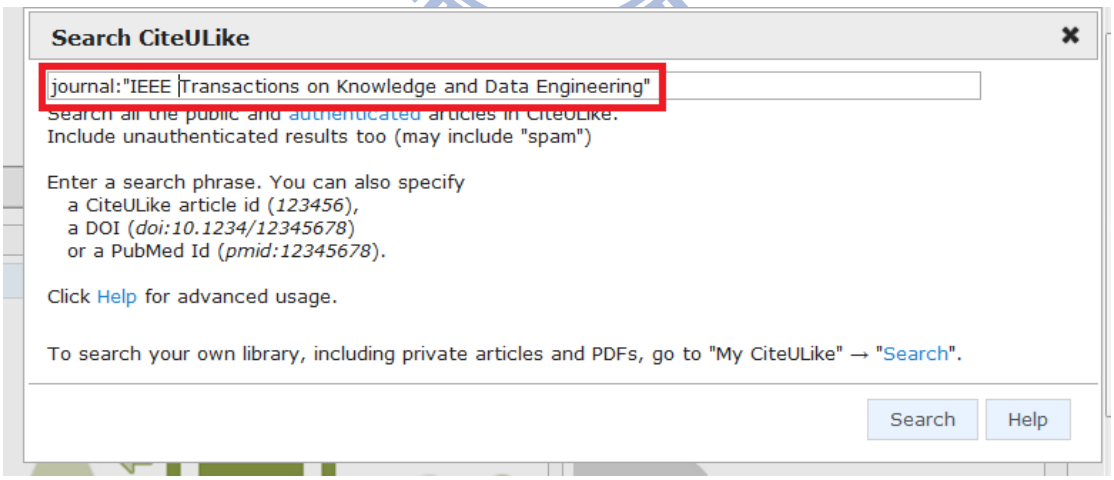


圖4-2：CiteULike期刊搜尋

搜尋之後會有許多的文章，其中一篇的文章格式如下圖 4-3，主要是抓取這

[2] http://www.citeulike.org/

篇文章的摘要部分以及標籤的部分。

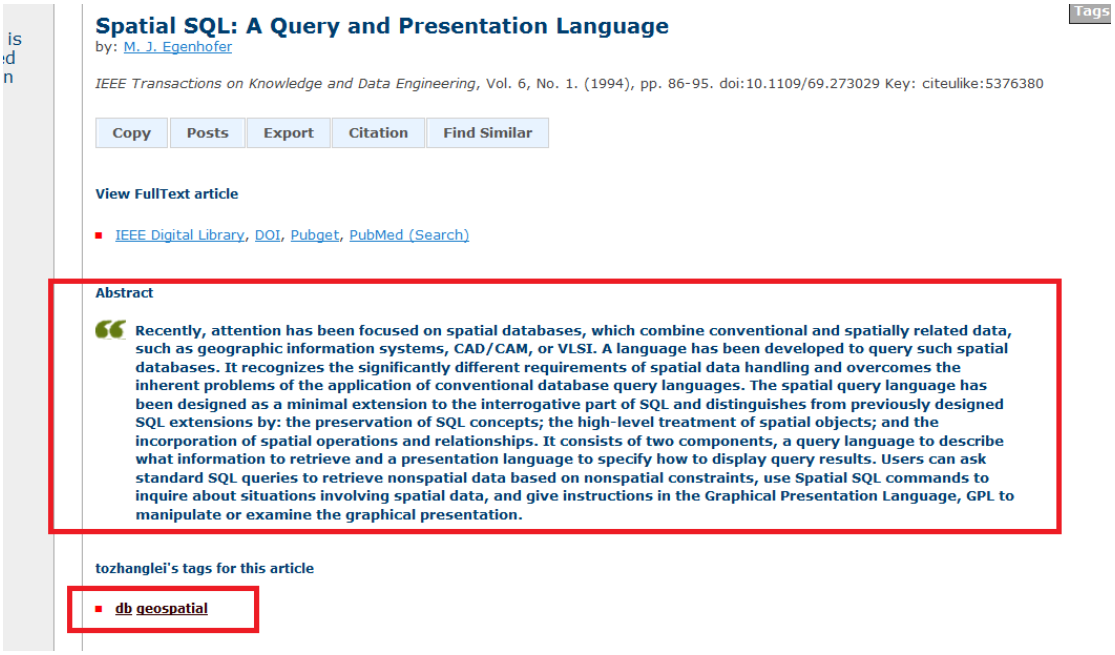


圖4-3：CiteULike期刊格式

我們收集了類別下許多的期刊，每個期刊下又有許多的論文，所以抓下來的論文資料集會如下圖 4-4 所示，最上層會是類別，中間層會有這類別下的期刊，每個期刊下有論文文章，然後每篇文章會有摘要和標籤。

本論文此部分目的在研究摘要和標籤要如何組合，會使得每篇論文分到類別的正確率最好。

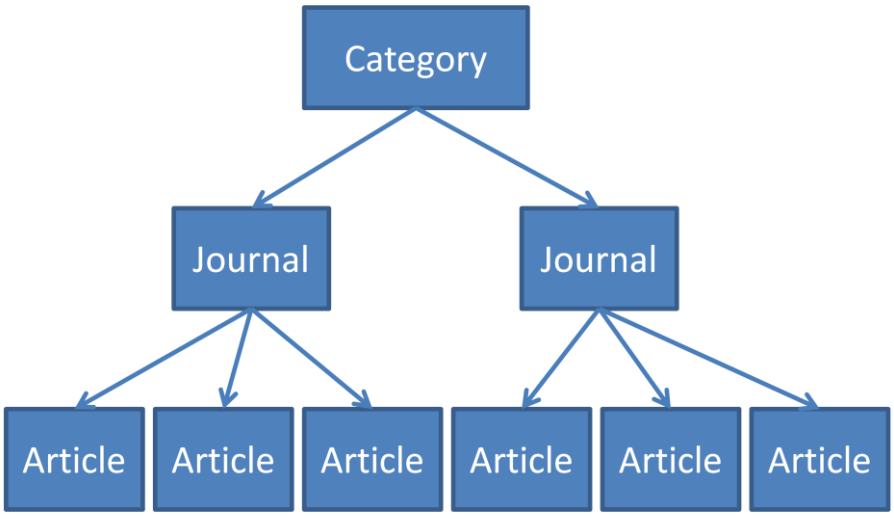


圖4-4：論文資料集架構

4.1.3　摘要標籤向量表示

(一)摘要標籤定義

一篇文章包含摘要和標籤兩個部分，標籤通常代表一篇文章的關鍵字或者屬性，所以理論上，標籤有助於文章的分類。以下討論摘要和標籤要如何以向量表示，分成下列四種組合方式：

1.只有摘要(Words Only)

字的集合為所有文章中摘要的字，一篇論文用摘要字的集合表示成一個向量，每一個向量數值表示一篇文章摘要的字出現在字的集合的次數，最後再對此向量作正規化(Normalize)的處理。

2.只有標籤(Tags Only)

與只有摘要(Words Only)狀況類似，只用標籤來表示文章，標籤的集合為出現在所有文章標籤中的字，一篇論文即可用此標籤的集合表示成一個向量，向量數值表示一篇文章標籤的字出現在標籤的集合的次數，最後再對此向量作正規化(Normalize)的處理。

3.摘要標籤個別比例(Words+Tags)

字的集合為所有出現在摘要中的字，標籤的集合為所有出現在標籤中的字，摘要和標籤有個別的向量，由個別的集合算出現次數，做正規化(Normalize)處理時，摘要和標籤向量依不同的比例組合。

4.標籤當成摘要字(Tags as Words)

結合字的集合和標籤的集合為一個新的集合，也就是對這兩個集合取聯集，標籤的權重為n，假設一篇論文摘要出現“big”一次，標籤也出現“big”，則 big 這個字在向量中權重為 1+n，最後一樣做正規化(Normalize)處理。

(二)摘要標籤例子

字的集合為：a b c d e

標籤的集合為：c f

某一篇文章的摘要：a b

某一篇文章的標籤：c f

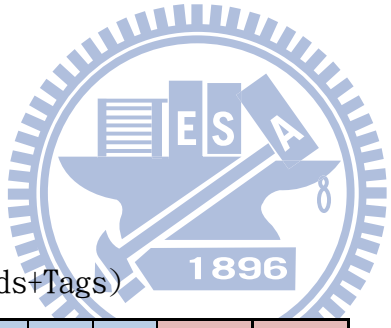
以下則為某一篇文章的向量表示方式：

1.只有摘要(Words Only)

	a	b	c	d	e
vector	0.5	0.5	0	0	0

2.只有標籤(Tags Only)

	c	f
vector	0.5	0.5



3.摘要標籤個別比例(Words+Tags)

	a	b	c	d	e	c‘	f
vector	0.4	0.4	0	0	0	0.1	0.1

Words+Tags 當比例為 8:2 的情況

4.標籤當成摘要字(Tags as Words)

	a	b	c	d	e	f
vector	0.1	0.1	0.4	0	0	0.4

當標籤權重 n 為 4

4.2 實驗資料

本論文使用的資料集共有三種，每種的資料集有其不同的特性，其類別與文章數也不太相同，以下介紹這三種資料集。

1. CiteUlike 論文資料集

此資料集如之前第三章所介紹的，主要是收集期刊上的論文摘要，還有使用者給的標籤(Tag)，來當作分類的依據；本論文收集了三個類別 Databases、Graphics、Programming Languages，這三個類別的期刊資訊如下表 4-1 所示。

表 4-1：論文資料集 A

DATABASES		GRAPHICS		PROGRAMMINGLANGUAGES	
CSDA	813	CG	285	JFP	116
DKE	249	CGA	88	MP	381
SIGMOD	43	CGF	143	SCP	294
TKDE	169	TOG	32	SIGPLAN	491
TODS	15	TVCG	193	TOPLAS	82
Total	1289	Total	741	Total	1364

由於某些期刊的內容與所屬的類別不符，因此去掉一些不相似期刊，如下表 4-2 所示。

表 4-2：論文資料集 B

DATABASES		GRAPHICS		PROGRAMMINGLANGUAGES	
CSDA	813	CG	285	JFP	116
		CGA	88	SCP	294
		CGF	143	SIGPLAN	491
		TOG	32	TOPLAS	82
		TVCG	193		
Total	813	Total	741	Total	983

2. 20 Newsgroups

20 Newsgroups 是一個常被用在文件分群或分類的資料集，此資料集從網路新聞討論串取出，包含了將近 20000 篇的文章，共有 20 個類別，包含電腦類、運動類、政治類等類別，本論文取出 comp、talk、rec、sci 這 4 個類別來做分析，共分成 3 種情況來做實驗，每個類別都是 1000 篇文章：

(1) 20newsgroups 資料集 A

包含 comp.sys.ibm.pc.hardware 和 comp.sys.mac.hardware。

(2) 20newsgroups 資料集 B

包含 talk.politics.guns 和 talk.politics.misc。

(3) 20newsgroups 資料集 C

包含 comp.graphics、rec.autos、sci.crypt 和 talk.politics.guns。

3. Reuters

Reuters-21578 也是一個很常用來做文件分類或分群的一個資料集，這個資料集共有 21578 篇文章，這些文章是用人工分類，全部共有 135 個類別；因為所有文章包含的類別太多，因此取出類別文章數最高的 10 類來做分析，下表 4-3 表示此 10 類的資訊，該種實驗方式也廣泛的被其他研究學者所採用。

表 4-3：Reuters 資料集

ACQ	2131
corn	207
crude	510
earn	3753
grain	528
interest	389
money-fx	601
ship	276
trade	449
wheat	264
total	9108

4.3 實驗步驟

4.3.1 前處理

(1)Port Stemming

我們使用 Stemming 的目的是為了將意思相同但型態不同的字轉化成一樣的字，例如，movies 會變化成 movi、movie 也會變化成 movi，如此一來，movies 和 movie 會轉換成同一個字；在計算特徵向量時，他們本來就是相同意思，應該用同一個字來表現。我們所採用的是 Porter 的演算法，這樣能將許多字母的變化型去除掉，減少許多字的變化形態，相同意義的字，皆用同一個字來代表。

(2)Stop Words

在文章之中，會有許多的介系詞、語助詞之類的文字，像是 a、the、an 等等，也就是 Stop Words，這些字大多不具意義，且會造成分群的一些雜訊(Noise)，因此我們將一些常見的 Stop words 建成一個表格，在將資料做處理的時候，會參照這個表，將不必要的 Stop words 去除掉。

(3)變小寫、去除特殊符號

在文章中，作者使用的字可能會有大小寫不同的情況，電腦會將大小寫看成是不同的字，因此我們將其所有的文字都改成小寫。文章中還會有許多雜訊的特殊符號，例如標點符號、羅馬符號等等，這些對於文章的分類、分群是沒什麼幫助的，因此我們會把這些符號刪除。

4.3.2 標籤分析實驗

在標籤分析實驗中，本論文用 clustering 的演算法來進行實驗，如第三章所提到的四種組合，words only、tags only、words+tags 和 tags as words。利用這幾組組合來進行實驗：

- words only 和 tags only

  就如同一般分群的情況，將其建出來的向量直接分群。

- words+tags 比例法

  我們分別依照 word:tag 比例從 1:9、2:8 一直到 9:1 進行實驗。

- tags as words

  設定標籤的權重，我們設定此值為 1 到 20。

實驗比較用到的方法有 PLSA[23]、LDA[29]、K-Means、MM-LDA[30]，這些方法皆為非監督式演算法。本論文也放入了 Constrained-PLSA 來與這些非監督式方法做比較，這邊標記的資訊只有 1%，比較加入些許標記資訊效果是否比上述的非監督式方法好。

4.3.3 Semi-Supervised Learning 實驗

Semi-Supervised Learning 實驗中，我們取標記資料的方式是依照不同比例來分析系統效能分成 1%到 5%，每次取的標記資料都是從每個類別隨機取出。假設現在的資料集有 2 類別，各有 1000 篇文章，在 1%的情況下我們從這兩個類別個別隨機取出 10 篇來當作標記資料；另外考慮到公平性與客觀性，我們隨機取 10 次做實驗，並將 10 次結果的平均做為實驗結果；主要是為了避免挑到比較好的或比較差的標記資料。此部分的實驗所用到的比較方法有 Constrained-PLSA、Constrained-KMeans[10]、tSVM[7]、Graph-based Semi-supervised[9]。

4.3.4 Constrained-PLSA 流程圖

下圖 4-5 為本論文所提出方法 Constrained-PLSA 之實驗流程圖。

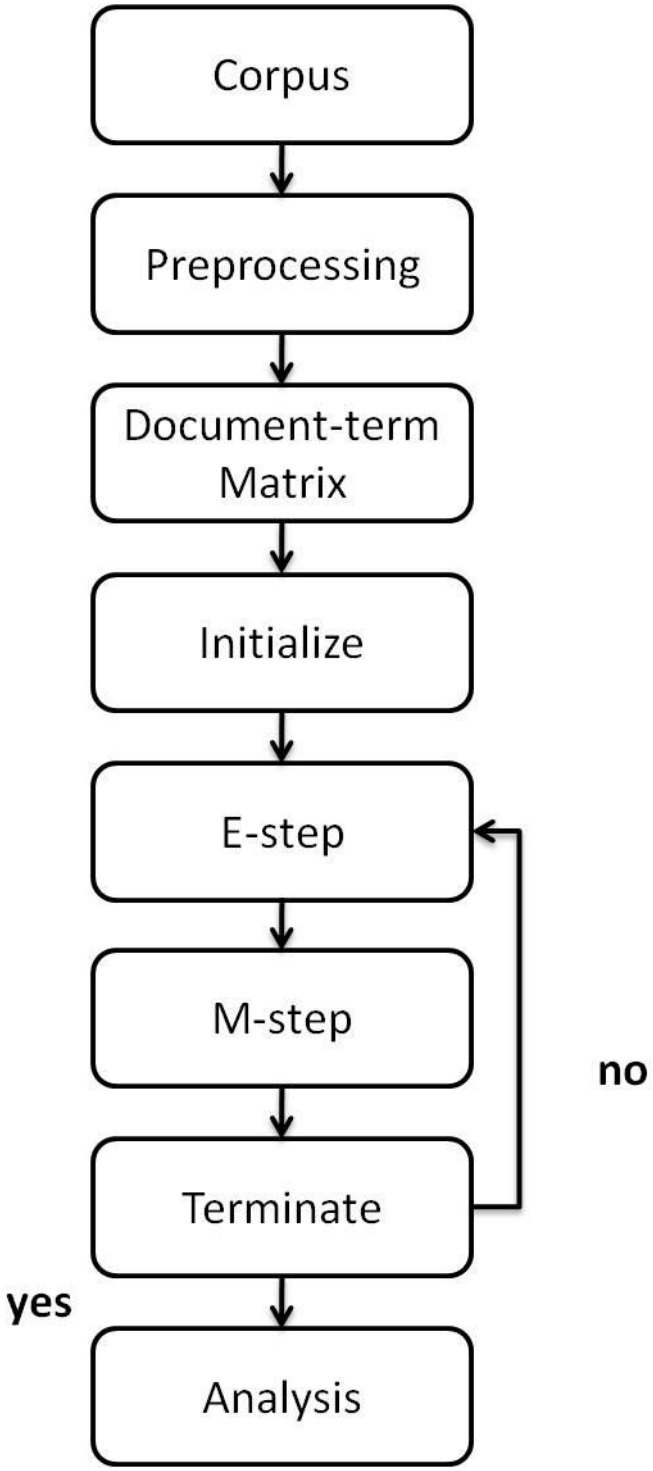


圖4-5：Constrained-PLSA流程

4.4 效能計算方式

4.4.1 F1 cluster evaluation measure

非監督式的部分，本論文採用 F1 cluster evaluation measure[30]來評估分群效能。當系統在做分群時，系統並不會知道哪個群是哪個類別，系統只會將所有的資料分成指定的群數，所以計算效能時，是將分群的結果和真實的類別作比較，比較方式可分成四種情況：

1. True Positives(TP)：

   系統將兩篇文章分在同一群，而這兩篇文章實際也是在同個類別內。

2. False Positives(FP)：

   系統將兩篇文章分在同一群，但是這兩篇文章實際不是在同個類別內。

3. True Negatives(TN)：

   系統將兩篇文章分在不同群，而這兩篇文章實際也不在同個類別內。

4. False Negatives(FN)：

   系統將兩篇文章分在不同群，但是這兩篇文章實際是在同個類別內。

F1 cluster evaluation measure 的算法定義如下：

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * precision * recall}{precision + recall}$$

4.4.2 Macro-average F-measure

半監督式的部分，本論文採用 Macro-average F-measure 來評估分類後的結果好壞，而不是利用 accuracy 來分析，這是因為我們的資料集中，有些資料集每個類別的文章分布是不平均的，如 Reuters 這個資料集，在文章不平均的情況下，可能會發生不公平的問題，例如某資料集假設 A 類別是 800 篇、B 類別是 200 篇，系統都將文章分類到 A 類別則都會有 80%的效果，但其實此系統 B 類別全錯，因此利用 Macro-average F-measure 來分析是比較公平合理的，他會分析每個類別分布的情況。

Macro-average F-measure 是要計算每個類別的 F-measure 然後再來取平均，這裡的定義與 F1 cluster evaluation measure 是不相同的。

1. True Positives(TP)：

   系統將一篇正確類別的文章分到正確類別。

2. False Positives(FP)：

   系統將一篇錯誤類別的文章分到正確類別。

3. True Negatives(TN)：

   系統將一篇錯誤類別的文章分到錯誤類別。

4. False Negatives(FN)：

   系統將一篇正確類別的文章分到錯誤類別。

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F - measure = \frac{2 * precision * recall}{precision + recall}$$

$$Macro - average\ F - measure = \frac{\sum_{i=1}^{N} F - measure_i}{N}$$

4.5 實驗結果

4.5.1 標籤分析實驗結果

下表 4-4、表 4-5 和表 4-6 皆是論文資料集 A 所做出的實驗

表 4-4：論文資料集 A tags only、words only

	PLSA	K-means	LDA	Constrained-PLSA
Words only	0.4762	0.4842	0.504	0.5879
Tags only	0.5155	0.5013	0.5154	0.5766

表 4-5：論文資料集 A words+tags

比例	PLSA	K-means	MMLDA	Constrained-PLSA
9:1	0.465	0.4982	0.5126	0.5794
8:2	0.507	0.5128	0.5763	0.6173
7:3	0.5351	0.5236	0.5702	0.6026
6:4	0.5185	0.5219	0.5517	0.6139
5:5	0.5034	0.5116	0.5208	0.5478
4:6	0.4967	0.5204	0.4798	0.5201
3:7	0.4933	0.5232	0.462	0.5274
2:8	0.4935	0.5022	0.4907	0.5093
1:9	0.4991	0.4945	0.406	0.4913

表 4-6：論文資料集 A tags as word

Tag weight	PLSA	K-means	LDA	Constrained-PLSA
1	0.5094	0.4887	0.5573	0.6091
5	0.5707	0.4783	0.5719	0.6154
10	0.5622	0.4957	0.5605	0.6163
15	0.5831	0.4994	0.5631	0.6282
18	0.5933	0.5	0.5874	0.6236
20	0.5503	0.5002	0.5043	0.5947

下表 4-7、表 4-8 和表 4-9 皆是論文資料集 B 所做出的實驗

表 4-7：論文資料集 B tags only、words only

	PLSA	Kmeans	LDA	Constrained-PLSA
Words only	0.8108	0.4754	0.8212	0.9265
Tags only	0.5793	0.5111	0.5456	0.6416

表 4-8：論文資料集 B words+tags

比例	PLSA	Kmeans	MMLDA	Constrained-PLSA
9:1	0.9428	0.5534	0.6231	0.9432
8:2	0.9057	0.5828	0.619	0.9268
7:3	0.6308	0.4793	0.9052	0.9069
6:4	0.5831	0.4827	0.8752	0.8854
5:5	0.5481	0.5043	0.8093	0.6508
4:6	0.507	0.5042	0.7298	0.5811
3:7	0.4886	0.503	0.5395	0.5494
2:8	0.4799	0.5025	0.5589	0.5159
1:9	0.4727	0.5026	0.5347	0.4987

表 4-9：論文資料集 B tags as word

Tag weight	PLSA	Kmeans	LDA	Constrained-PLSA
1	0.9435	0.5594	0.9346	0.9467
2	0.9428	0.5707	0.9316	0.9453
3	0.9345	0.4802	0.9254	0.9444
5	0.9316	0.4787	0.8911	0.9363
10	0.9157	0.4843	0.8449	0.9206

4.5.2 Semi-Supervised Learning 實驗結果

下表 4-10 和圖 4-6 為 20newsgroups 資料集 A 數據結果和曲線圖，20newsgroups 資料集 A 是取 comp.sys.ibm.pc.hardware 和 comp.sys.mac.hardware 這兩個類別。

表 4-10：20newsgroups 資料集 A 數據結果

	Constrained -PLSA	tSVM	Graph-based Semi-supervised	Constrained -KMeans
1%	0.6983	0.5386	0.6029	0.4922
2%	0.7380	0.5727	0.6332	0.5239
3%	0.8052	0.5926	0.6665	0.5354
4%	0.8172	0.6494	0.7119	0.5408
5%	0.8249	0.6774	0.7324	0.5524

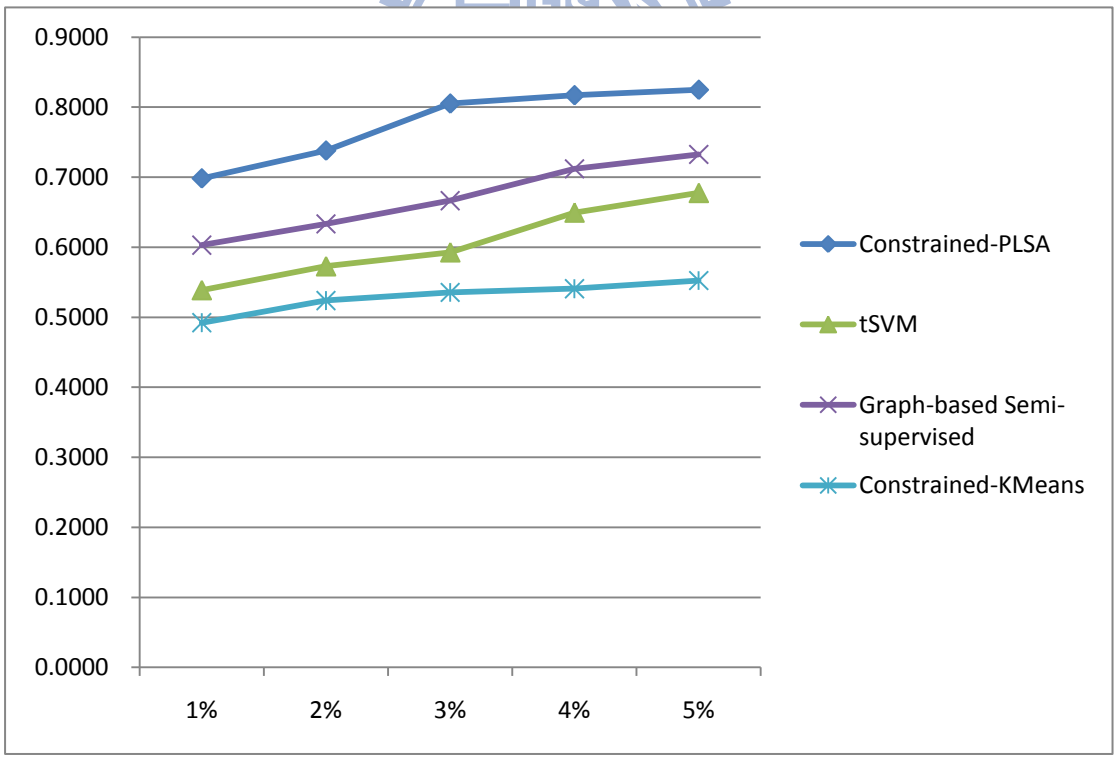


圖4-6：20newsgroups資料集A曲線圖

下表 4-11 和圖 4-7 為 20newsgroups 資料集 B 數據結果和曲線圖，20newsgroups 資料集 B 是取 talk.politics.guns 和 talk.politics.misc 這兩個類別。

表 4-11：20newsgroups 資料集 B 數據結果

	Constrained -PLSA	tSVM	Graph-based Semi-supervised	Constrained -KMeans
1%	0.6956	0.5325	0.5232	0.4630
2%	0.7781	0.5658	0.6448	0.4916
3%	0.7918	0.5996	0.6583	0.4976
4%	0.8116	0.6229	0.7037	0.5078
5%	0.8122	0.6334	0.7111	0.5161

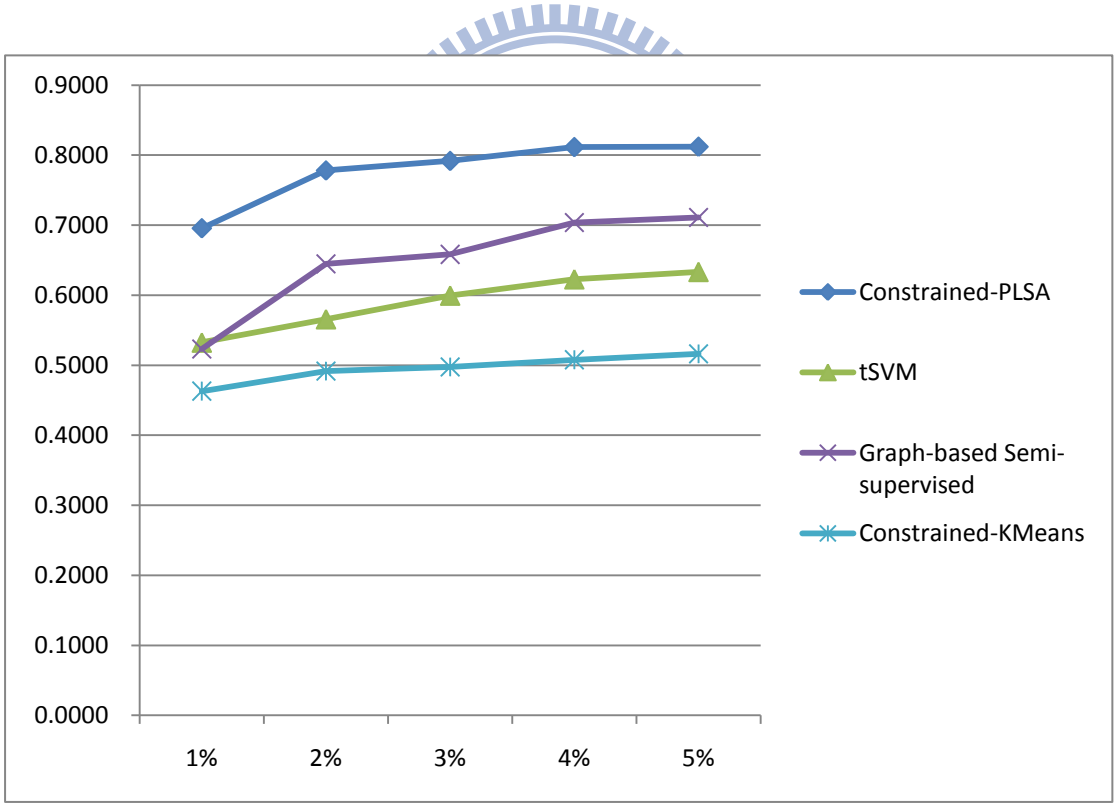


圖4-7：20newsgroups資料集B曲線圖

下表 4-12 和圖 4-8 為 20newsgroups 資料集 C 數據結果和曲線圖，20newsgroups 資料集 C 是取 comp.graphics、rec.autos、sci.crypt 和 talk.politics.guns 這四個類別。

表 4-12：20newsgroups 資料集 C 數據結果

	Constrained -PLSA	tSVM	Graph-based Semi-supervised	Constrained -KMeans
1%	0.9356	0.5559	0.3578	0.2813
2%	0.9371	0.7084	0.3808	0.2886
3%	0.9376	0.7431	0.4042	0.3141
4%	0.9378	0.7803	0.4282	0.3421
5%	0.9395	0.8141	0.4311	0.3491

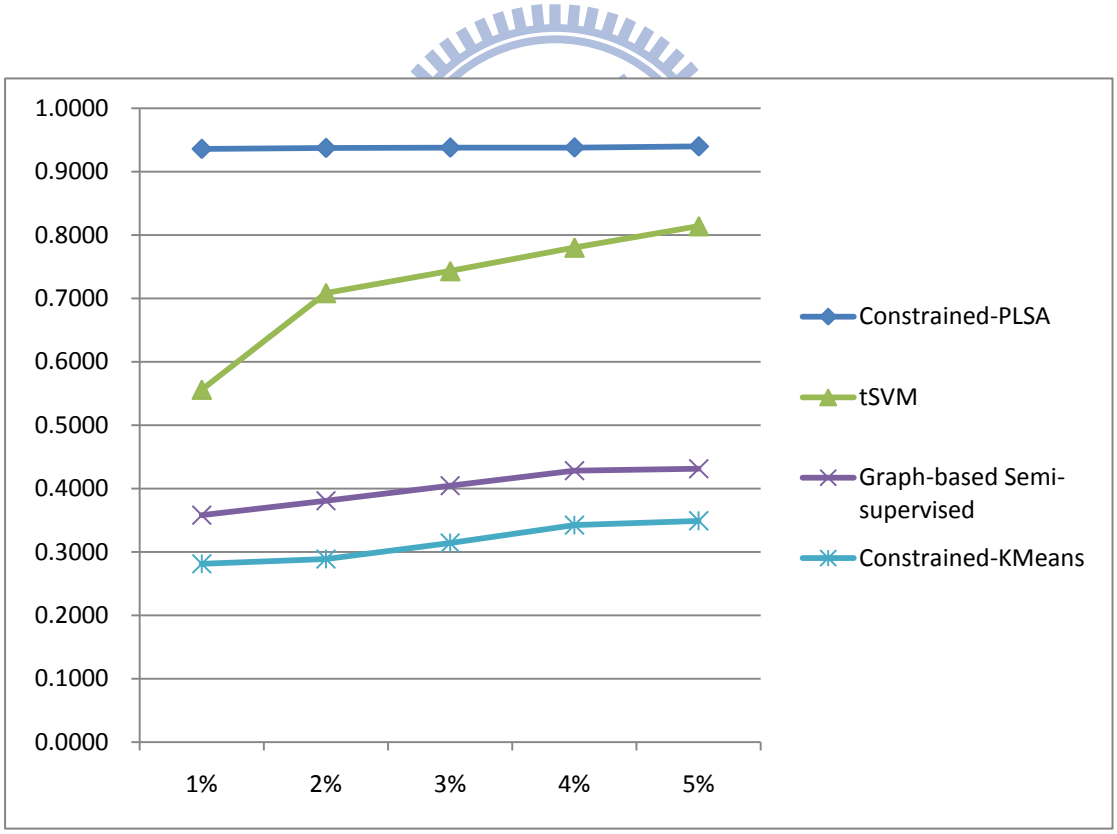


圖4-8：20newsgroups資料集C曲線圖

下表 4-13 和圖 4-9 為 Reuters 資料集數據結果和曲線圖，Reuters 資料集則取前 10 大的群做實驗。

表 4-13：Reuters 資料集數據結果

	Constrained -PLSA	tSVM	Graph-based Semi-supervised	Constrained -KMeans
1%	0.5159	0.4374	0.1177	0.2493
2%	0.5351	0.5348	0.1331	0.3095
3%	0.5389	0.5532	0.1492	0.3136
4%	0.5484	0.5930	0.1545	0.3447
5%	0.5707	0.6111	0.1587	0.3567

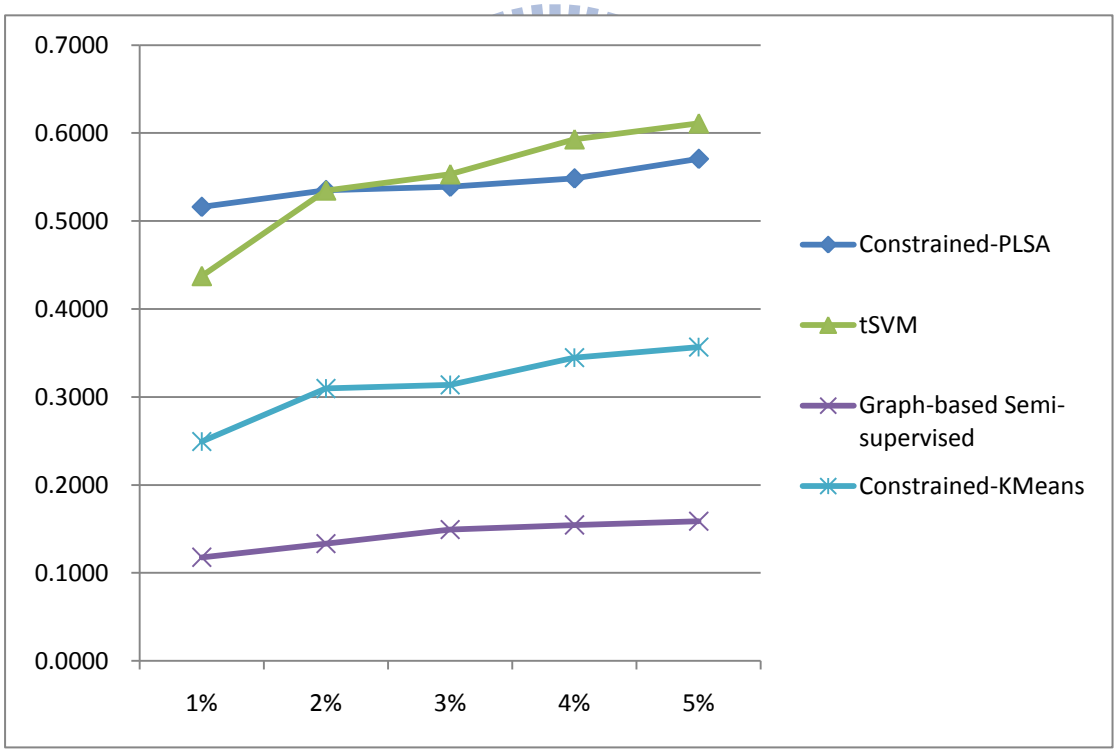


圖4-9：Reuters資料集曲線圖

4.6 實驗討論

4.6.1 標籤分析討論

在標籤分析實驗中，在論文資料集 A 的情況下 Words Only 和 Tags Only 的情況下效果是不太好的，如表 4-4 所示，單看摘要或標籤並不能有效的分群。Words+Tags 和 Tags as words 的狀況下都能幫助分群的效果，如表 4-5 和表 4-6 所示，效果有些許的提升；Tags as words 的在權重約 18 的時候效果最好，當權重超過 18 之後，效果反而會下降，這是因為標籤的權重太高了，所以摘要和標籤要適當的給予權重。在論文資料集 B 的情況下，因為將一些關聯性較低的期刊移除，所以 Words Only 的實驗效果不錯，如表 4-7 所示，但 Tags Only 因為每篇文章的標籤資訊量不多，所以效果不太好。在 Tags as words 的實驗，標籤的權重只要 1 就能讓效能提升到 94%，如表 4-9 所示，實驗最後的結果也是 Tags as words 能夠使標籤達到最好的效果。

Constrained-PLSA 只加入了 1%的標記資訊，就比其他分群法效果來的好，從論文資料集 A 和論文資料集 B 的數據中可以看出，論文資料集 A 的資料集較難分，加入一點標記資訊，就可以大幅提升效果。在論文資料集 B 中，分群演算法效果已經不錯，因此Constrained-PLSA在標記資訊的輔助下只有小幅度的提升。可看出在資料集較難分的情況下，加入些許標記資訊是可以有效幫助分群的效果提升。

4.6.2 Semi-Supervised Learning 實驗討論

在 Semi-Supervised Learning 實驗中，本論文所提出的 Constrained-PLSA 在種子數 1%至 5%的情況下，大部分都有穩定又準確的效能。在 20newsgroups 資料集 A 和 20newsgroups 資料集 B 中，兩個類別的情況下，可由表 4-10 和表 4-11 看出，比效能第二的演算法 Graph-based Semi-supervised 好 10%左右，更是比 tSVM 和 Constrained-KMeans 好 15%以上，在這資料集中，從圖 4-6 和圖 4-7 可清楚看出本論文之方法與其他方法的差距，Constrained-PLSA 明顯優勢於其

他演算法。

20newsgroups 資料集 C，是四個類別的情況，由表 4-12 可看到，本論文之方法效能可以達到 93%，代表 Constrained-PLSA 可以達到非常好的分群效果。Graph-based Semi-supervised 當類別數多的情況下，效能會越往下降，掉到只剩 40%左右。從圖 4-8 圖中可以看出，除了 tSVM 效能種子數 5%時可以達到 80%，其他皆低於 80%，Constrained-KMeans 和 Graph-based Semi-supervised 更是低於 50%，由此可看出在 20newsgroups 資料集 C 實驗，本論文之 Constrained-PLSA 效果相當好。

在 Reuters 資料集中，由表 4-13 可看出本論文之方法還是有相當好的效能，只有在 3%至 5%的情況下，略輸 tSVM5%以內，這是因為此資料集每群的文章數差異較大，Constrained-PLSA 較不適合分群此資料集；由圖 4-9 可看出 Constrained-PLSA 效能還是比 Constrained-KMeans 和 Graph-based Semi-supervised 高出許多，本論文之方法在大部分的情況下效能還是很好。

第五章、結果與展望

5.1 研究總結

標籤分析中，本論文研究發現在我們設定的四種向量組合中，利用標籤(Tag)來輔助分群，Tags as words 是最好組合的方式，優於其他三種組合方式。Words only 和 Tags only 特徵數都較少，無法達到好的效果，Words+tags 將標籤和摘要的比例分開，沒有集中，效果比 Tags as words 差一點。在這部分實驗中可看出 Constrained-PLSA 加入些許標記的資訊，很明顯的優於其他分群法，些許的標記資訊，可以有效的提升效能。

在 Semi-supervised Leaning 實驗中，本論文所提出的方法，將背景知識(Background Knowledge)，加入 Constrained-PLSA 演算法，能有效的提升效能且穩定的效果。在實驗結果中，在兩個類別和四個類別的情況下，Constrained-PLSA 都能達到很好的效果，與其它方法做比較，大約都高 10%左右的效能。Constrained-PLSA 只給予些許的標記(Labeled)資訊，使分群的效果達到不錯的水準，讓系統能夠有效的應用於文件管理應用、資訊擷取等應用上。

5.2 未來研究

本論文之 Constrained-PLSA 在大部分的情況下都有不錯的效能，但在 Reuters 實驗中發現種子數 3%至 5%的情況下略輸 tSVM，但效果還是有一定的水準，未來可以研究 Constrained-PLSA，要怎麼修改演算法，讓本論文之方法能夠在這個部分也比 tSVM 效能好。

參考文獻

[1] T. Joachims, "Text categorization with suport vector machines: Learning with many relevant features," in Proceedings of the 10th European Conference on Machine Learning. London, UK: Springer-Verlag, 1998, pp. 137–142. [Online]. Available: http://portal.acm.org/citation.cfm?id=645326.649721

[2] A. McCallum and K. Nigam, "A comparison of event models for naive bayes text classification," in IN AAAI-98 WORKSHOP ON LEARNING FOR TEXT CATEGORIZATION. AAAI Press, 1998, pp. 41–48.

[3] R. E. Schapire and Y. Singer, "Boostexter: A boosting-based system for text categorization," Machine Learning, vol. 39, no. 2/3, pp. 135–168, 2000.

[4] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in Proceedings of the eleventh annual conference on Computational learning theory, ser. COLT' 98. New York, NY, USA: ACM, 1998, pp. 92–100. [Online]. Available: http://doi.acm.org/10.1145/279943.279962

[5] W. Wang and Z.-H. Zhou, "A new analysis of co-training," in Proceedings of the 27th International Conference on Machine Learning (ICML-10), J. F¨urnkranz and T. Joachims, Eds. Haifa, Israel: Omnipress, June 2010, pp. 1135–1142. [Online]. Available: http://www.icml2010.org/papers/275.pdf

[6] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," Mach. Learn., vol. 39, pp. 103–134, May 2000. [Online]. Available: http://portal.acm.org/citation.cfm?id=347709.347724

[7] T. Joachims, "Transductive inference for text classification using support vector machines," in Proceedings of the Sixteenth International Conference on Machine Learning, ser. ICML '99. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1999, pp. 200–209. [Online]. Available: http://portal.acm.org/citation.cfm?id=645528.657646

[8] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 19–26. [Online]. Available: http://portal.acm.org/citation.cfm?id=645530.757779

[9] A. B. Goldberg and X. Zhu, "Seeing stars when there aren't many stars: graph-based semi-supervised learning for sentiment categorization," in Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing, ser. TextGraphs-1. Stroudsburg, PA, USA: Association for Computational Linguistics, 2006, pp. 45–52. [Online]. Available: http://portal.acm.org/citation.cfm?id=1654758.1654769

[10] K. Wagstaff, C. Cardie, S. Rogers, and S. Schr¨odl, "Constrained k-means clustering with background knowledge," in Proceedings of the Eighteenth International Conference on Machine Learning, ser. ICML '01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 577–584. [Online]. Available: http://portal.acm.org/citation.cfm?id=645530.655669

[11] S. Basu, A. Banerjee, and R. J. Mooney, "Semi-supervised clustering by seeding," in Proceedings of the Nineteenth International Conference on Machine Learning, ser. ICML '02. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2002, pp. 27–34. [Online]. Available: http://portal.acm.org/citation.cfm?id=645531.656012

[12] S. Basu, M. Bilenko, and R. J. Mooney, "A probabilistic framework for semi-supervised clustering," in Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ser. KDD '04. New York, NY, USA: ACM, 2004, pp. 59–68. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014062

[13] K. Zhou, X. Gui-Rong, Q. Yang, and Y. Yu, "Learning with positive and unlabeled examples using topicsensitive plsa," IEEE Trans. on Knowl. and Data Eng., vol. 22, pp. 46–58, January 2010. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2009.56

[14] W. Li and A. McCallum, "Semi-supervised sequence modeling with syntactic topic models," in Proceedings of the 20th national conference on Artificial intelligence - Volume 2. AAAI Press, 2005, pp. 813–818. [Online]. Available: http://portal.acm.org/citation.cfm?id=1619410.1619463

[15] A. Carlson, J. Betteridge, R. C. Wang, E. R. Hruschka, Jr., and T. M. Mitchell, "Coupled semi-supervised learning for information extraction," in Proceedings of the third ACM international conference on Web search and data mining, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 101–110. [Online]. Available: http://doi.acm.org/10.1145/1718487.1718501

[16] S. X. Yu and J. Shi, "Grouping with bias," in NIPS, 2001, pp. 1327–1334.
[17] R. Nock and F. Nielsen, "Grouping with bias revisited," in Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition, ser. CVPR'04. Washington, DC, USA: IEEE Computer Society, 2004, pp. 460–465. [Online]. Available: http://portal.acm.org/citation.cfm?id=1896300.1896368
[18] T. Joachims, "Transductive learning via spectral graph partitioning," in Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), 2003, pp. 290–297.
[19] X. Zhu, "Semi-supervised learning literature survey," Computer Sciences, University of Wisconsin-Madison, Tech. Rep. 1530, 2005.
[20] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol. 22, no. 8, pp. 888–905, 2000.
[21] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in Advances in Neural Information Processing Systems 14. MIT Press, 2001, pp. 849–856.
[22] T. Hofmann, "Probabilistic latent semantic analysis," in Proc. of Uncertainty in Artificial Intelligence, UAI'99, 1999. [Online]. Available: citeseer.ist.psu.edu/hofmann99probabilistic.html
[23] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," Mach. Learn., vol. 42, no. 1-2, pp. 177–196, 2001.
[24] T. Hofmann, J. Puzicha, and M. I. Jordan, "Learning from dyadic data," in Proceedings of the 1998 conference on Advances in neural information processing systems II. Cambridge, MA, USA: MIT Press, 1999, pp. 466–472.
[25] S. Zhong, "Semi-supervised model-based document clustering: A comparative study," Mach. Learn., vol. 65, pp. 3–29, October 2006. [Online]. Available: http://portal.acm.org/citation.cfm?id=1164582.1164590
[26] T. Finley and T. Joachims, "Supervised clustering with support vector machines," in Proceedings of the 22nd international conference on Machine learning, ser. ICML '05. New York, NY, USA: ACM, 2005, pp. 217–224. [Online]. Available: http://doi.acm.org/10.1145/1102351.1102379
[27] J. Wang, S. Wu, H. Q. Vu, and G. Li, "Text document clustering with metric learning," in Proceeding of the 33rd international ACM SIGIR

conference on Research and development in information retrieval, ser. SIGIR '10. New York, NY, USA: ACM, 2010, pp. 783–784. [Online]. Available: http://doi.acm.org/10.1145/1835449.1835614

[28] 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. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.133.4884

[29] D.M.Blei, A.Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation" in Journal of Machine Learning Research, 2003, pp. 993-1002

[30] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina, "Clustering the Tagged Web" in Second ACM International Conference on Web Search and Data Mining (WSFM), 2009