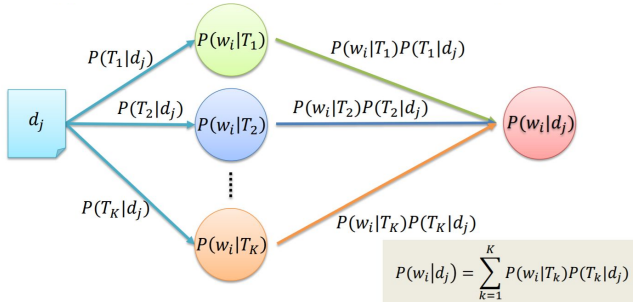
Information Retrieval and Applications

HW03-PLSA model with EM algorithm

四資工四甲B10515047陳牧凡

此次作業運用到的MODEL是PLSA model，跟前兩個作業最大的差別在於這個model有靠慮到語言中會有同義詞或是同詞異義的問題，例如同樣"java"一詞在講人文地理時可能是一座島嶼、或是一種咖啡豆，但在電腦資訊主題裡面的話就很有可能是指程式語言的那個java。為了解決這個問題我們會自訂義一個變數"主題(Topic)"，透過機率去計算這篇文章出現在該主題的機率、以及這些字詞出現在該主題的機率，交互去比對之後算出他們的文章相似度，如下圖。



那麼訓練其topic的方式是採用EM algorithm來進行計算:

- E-step

$$P(T_k|w_i, d_j) = \frac{P(w_i|T_k)P(T_k|d_j)}{\sum_{k=1}^{K} P(w_i|T_k)P(T_k|d_j)}$$

- M-step

$$P(w_i|T_k) = \frac{\sum_{d_j \in \mathbf{D}} c(w_i, d_j)P(T_k|w_i, d_j)}{\sum_{i'=1}^{|V|} \sum_{d_j \in \mathbf{D}} c(w_{i'}, d_j)P(T_k|w_{i'}, d_j)}$$

$$P(T_k|d_j) = \frac{\sum_{i=1}^{|V|} c(w_i, d_j)P(T_k|w_i, d_j)}{\sum_{i'=1}^{|V|} c(w_{i'}, d_j)} = \frac{\sum_{i=1}^{|V|} c(w_i, d_j)P(T_k|w_i, d_j)}{|d_j|}$$

透過不斷EM training迭代的方式來得到一組較佳的P(T|d)和P(w|T)，進而計算文章相似度。

參考資料:

Chen, L.C., Chen, D.R., Yeh K.H.and Wu, C.C. (2015), 'Using the Semantic Models to Analyze the Online Blog Posts', Journal of Information Management, Vol. 22, No. 3, pp. 273-316.